HIV protein sequence signatures for crosstalk with host proteins

A Thesis

Submitted to the Faculty

of

Drexel University

by

Mahdi Sarmady

in partial fulfillment of the

requirements for the degree

of

Doctor of Philosophy

September 2010

Dedications

To my parents

and

in loving memory of my grandfather

Acknowledgements

As with all things in my life, I would first like to thank my family, specially my parents, Mohammad Hossein and Parvin, for giving me solid roots from which I can grow, for showing me the value of education, and for their unconditional support and encouragement by which I could freely pursue my studies.

I would like to deeply thank my advisor, Prof. Aydin Tozeren, for his guidance, patience, help, and support during the years of my doctoral studies. I have been very privileged and proud to have him as my advisor. Without his support, this thesis would have not been possible. He has always been there tirelessly for me at every step of the way and his moral and academic supports were always inspiring in times of distress.

I extend my sincere gratitude to Prof. Kambiz Pourrezaei who has been always supportive and encouraging from my first day at Drexel.

I would like to thank the rest of my thesis committee: Prof. Heinrich Roder, Prof. Afshin Daryoush, Dr. Andres Kriete, and Dr. Ahmet Sacan for their insightful comments.

I am grateful to current and past members of the Center for Integrated Bioinformatics at Drexel University: Noor Dawany, Will Dampier, Perry Evans, Yi Chuan Liu, and my other friends for their help and making my doctoral studies enjoyable.

Table of Contents

List of Tables

List of Figures

Abstract

HIV protein sequence signatures for crosstalk with host proteins

Mahdi Sarmady

Aydin Tozeren, PhD

The HIV virus targets the immune system cells and suppresses immunity. The topology and connectivity of the signalling networks in host cells infected with the HIV virus are altered and redirected toward the synthesis of the virus. HIV proteins interact with host cell DNA and proteins in modulating cell signalling and metabolic pathways. Recent experimental studies involving immunoprecipitation and other binding assays have already identified a large number of host proteins as interacting with HIV virus proteins. Similarly, experiments with site-directed mutagenesis and HIV protein segments provided information on viral sequence sites potentially responsible for crosstalk with host proteins. Nevertheless, these experiments were not performed systematically and as a result much remains unknown about the HIV sequence hotspots for binding to host proteins.

My Ph.D. thesis focuses on the identification of HIV sequence hotspots, identities of their target proteins hotspots are used as binding interfaces, and the identities of host proteins outcompeted by viral proteins in these binding interactions. For this purpose I use bioinformatics databases containing large numbers of copies of viral sequences, previously annotated HIV-host protein interactions, and the host protein interactome.

The large-scale datasets on sequences and interactomes are integrated with motif discovery, statistical enrichment, and network construction tools in a computer code to reveal information on the details of binding interactions between HIV and host proteins.

This dissertation has produced a system wide portrayal of how HIV virus proteins interact with host hub proteins and the resulting changes in the host network. My work has also identified Nef sequence hotspots potentially initiating binding interactions with thirty or more host proteins. My findings are largely consistent with existing experimental data and suggest new experiments on binding interfaces as well as identify HIV virus sequence targets for drug discovery. In this thesis I have also illustrated the use of network analysis in constructing medically relevant cellular pathways such as the one depicting HIV virus interactions with host cell iron ion binding protein pathways. Taken together, my work produces bioinformatics and computational biology techniques specially designed to investigate crosstalk between a virus and the host.

Chapter 1: Introduction and background

1.1. Summary

Our view of biological science has changed drastically in the past decade with the development of high-throughput technologies. Traditionally, the biological data were obtained through lab work, which can be costly and laborious and leads to small data sizes. Emerging high-throughput technologies have paved the way to the production of large and multi-dimensional biological datasets. Examples of these high-throughput technologies include next-generation sequencing [1], yeast two hybrid method for protein binding interactions [2, 3] and protein chips [4].

The biological information produced by currently available high-throughput technologies is limited to the determination of sequences, global gene expression profiles and other assays used in pharmacology. Genes and proteins can be sequenced rather accurately and their expression levels can be measured with minimal noise. But the related molecular structure datasets for interpretation of high throughput data are not yet comprehensive; they are limited by the pace of protein structure determining methods which are still far behind other biological measurement techniques. Same argument holds for protein-protein interactions (PPIs). Current methods of discovery of binding interactions between proteins do not yield insight into the details of the interaction sites and their structure. These methods [3-6] indicate only that two proteins interact, but do not give information about structural and molecular mechanisms of the

interaction. Hence computational models and predictions of interaction site discovery can be very promising in the annotation of protein function by uncovering details of micromechanics of PPIs including binding sites.

Proteins orchestrate crucial activities and functions throughout the cellular life cycle. The functions of proteins are typically accomplished by interactions with other molecules including proteins, DNA, and RNA molecules. Interactions between proteins often involve formation of protein complexes [7], both in the form of transient structures or permanent structures. Tens of thousands of such interactions take place within a eukaryotic cell forming a network of interactions called the interactome [2].

Viral proteins alter host protein-protein interaction (PPI) networks by creating new interactions, modifying or destroying others. The resulting network topology favors excessive amounts of virus production in a stressed host cell network. Short linear peptide motifs common to both the virus and host provide the basis for host network modification [8]. A large percentage of transient PPIs occur due to the physical interaction of a short linear motif on one protein with a structural counter-motif on the partner protein [9, 10].

A number of recent studies focused on the identification of human proteins targeted by viral proteins. The list of virus-targeted host proteins is especially long in the case of the HIV-1 virus, arguably the most extensively studied virus at present. The motif/counter-motif pairs involved in virus-host PPI are yet to be discovered in most cases[11].

Annotation of motifs and counter-motifs on the interface of virus-host PPI will allow us to discover previously unknown virus-host interactions. Knowledge of the interacting components in virus-host PPI will impact the development of drugs that modulate or break bonds between virus and host proteins [12-14].

1.2. Protein - Protein Interaction Mechanisms

Protein functions are being carried out by its interactions with other proteins and molecules. Proteins participating in a PPI undergo conformation changes that facilitate the interaction. These structural changes may occur in a few regions of one protein opposing the interface with the counterpart [15]. In a recent study [16], Stein et al. investigated interactions of known 3D structure in the Protein Data Bank (PDB) [17] and showed that protein interactions can be divided into two major groups based on their contact interfaces: (i) Domain-domain interactions which involve the binding of two globular domains from different proteins which creates a large contact interface and (ii) Domain-peptide interactions in which a globular domain in one protein recognizes a short linear motif from another protein, thus making a rather small interface. Domain-peptide interactions are mostly observed in regulatory and signaling networks, which involve transient binding events. It has been shown that for signaling-pathway regulation and cell-compartment targeting there are most likely more linear motif instances than there are globular domains in the proteome [18]. Transient binding

interactions are much more difficult to be studied using biochemical experiments [19] but are important in such events as phosphorylation.

Tens of thousands of human protein-protein interactions have been reported in the literature. Multiple databases exist to store these experimentally validated PPI and update the databases as new interactions become available. A list of the publically available human PPI databases is shown in Table 1. Throughout this thesis, the Human Protein Reference Database (HPRD; [20]) is used as the reference human PPI database for the studies.

Table 1 Publically available human PPI databases details (source: [21])

Database	Number of human PPI	Number of Proteins	Downlad options
HPRD	36,617	9,427	Yes
BIND	6,621	3,887	Yes
DIP	1067	804	Yes
MINT	11,367	4,975	Yes
PDZBase	101	115	No
MIPS	346	405	Yes
IntAct	10244	4,614	Yes

1.3. Linear motifs

Linear motifs are short peptides of 3-10 amino acids length and are typically disordered to accommodate the more rigid interface on the opposing protein[22]. They are exposed to binding partners and have the ability to adapt to a variety of structural conformations. As mentioned in the previous section, domains in one protein recognize

a linear motif in their binding partner. Linear motifs involved in PPI can evolve in different organisms as they regularly arise and disappear by mutations hence presenting great adaptability to the interactome. The earliest definition of a short linear motif was presented by Tim Hunt in 1990 [23] :

"The sequences of many proteins contain short, conserved motifs that are involved in recognition and targeting activities, often separate from other functional properties of the molecule in which they occur. These motifs are linear, in the sense that three-dimensional organization is not required to bring distant segments of the molecule together to make the recognizable unit. The conservation of these motifs varies: some are highly conserved while others, for example, allow substitutions that retain only a certain pattern of charge across the motif."

Linear motifs demonstrate a particular sequence pattern which contains the key residues that can be recognised by the binding domain [24]. These key residues can be connected by variable residues, which guarantee proper spacing and provide the pattern of charge needed across the motif. *Regular expression patterns* are used to describe the consensus patterns of the linear motifs as they are recognizable by the computer and matching sequences can be matched against them rapidly. A common example is the Src-homology-3 (SH3) domain ligand. SH3 domain is shown to interact with Proline-rich regions of other proteins [25] . The regular expression for Class I SH3 ligand is [RK]..P..P [26] where `.' denotes arbitrary (wildcard) positions and brackets denote the set of possible residues for the position.

The Eukaryotic Linear Motif (ELM) database [27] contains more than 140 patterns for manually curated, experimentally verified consensus patterns of linear motifs. It uses regular expression in combination with logical filters to discriminate between likely true and false positives to improve the predictive value of the linear motifs of the database. The ELM database resource is the biggest database of its kind and was used throughout this research as the reference database of known linear motifs.

1.4. *De novo* Motif Discovery

The fact that 15- 40% of the interactions in the human proteome were estimated to be mediated by linear motifs [28] , suggests that hundreds of novel motif classes have to be discovered. Short length, low binding affinity, and extreme flexibility make linear motifs difficult for experimental analysis [8]. Emergence of proteome-scale interaction databases have paved the way for the development of fully computational tools for the discovery of *de novo* linear motifs associated with PPIs.

Multiple methods and approaches have been developed for *de novo* motif discovery using protein sets and protein interactome datasets [29-34]. The main hypothesis for all of these tools is that proteins sharing a common attribute, such as sub-cellular location, biological function or a common interaction partner, would share a feature that mediates that common attribute. This shared feature can be either a domain or a linear motif and in the absence of a shared domain, a linear motif could be the only common sequence feature which can be revealed by virtue of over-representation [35].

There are two main classes of *de novo* motif discovery tools: (i) Tools that discover pairs of correlated motifs within a set of interactions of the same type [32, 34], and (ii) Tools which look for over-represented linear motifs within a dataset of protein sequences with a common biological feature (i.e. shared interaction partner) [29, 30, 33, 35]. Discovery of correlated motifs on binding partners in an interactome subset reduces the discovery of motifs with no apparent function [34], but is not readily suitable to the present case of identifying motifs on large numbers of proteins interacting with the same protein. The dataset that can be used in the correlated approach should be highly balanced (symmetric) otherwise there will be too many false negatives in the results.

De novo motif discovery tools such as Discovery of Linear Motifs (DILIMOT) [33] are based on the TEIRESIAS [36] algorithm. TEIRESIAS is an algorithm for the discovery of patterns in biological sequences. This algorithm can reveal all patterns (presented by regular expressions) that appear in at least *m* (a user-defined number) sequences of a dataset. These patterns are maximal in the sense that they cannot be made by concatenating other patterns and they do not have full overlaps. TEIRESIAS works without enumerating the entire solution space and without using pairwise alignments which allows for enhanced performance [36]. To reduce the rate of false positive discovery, motif discovery tools including DILIMOT and SLiMFinder first detect homologous sequences within the input dataset using BLAST [37] to determine their evolutionary relationship and form unrelated protein clusters (UPC) , which are defined such that no protein in a UPC has a relationship with any protein in another UPC [30].

Then SLimFinder searches for motifs in all proteins and then weights results according to the evolutionary relationship for the proteins containing the motif. It also discovers patterns (motifs) with semi-conserved (wildcard) positions.

1.5. HIV - Human Protein Interactions

Viral proteins can interfere with host PPI networks by modifying and/or destroying the existent interactions or by creating new ones. The resulting topology allows for increased virus production within a stressed host cell network. The basis of these modifications is common short linear motifs that exist on both the viral and host proteins.[14, 38, 39]. As explained in the previous sections of the chapter, a relatively large percentage of transient PPIs are due to the physical interaction of a short linear motif on one protein with a structural counter-motif on the partner protein.

A Large number of experimental studies focused on the identification of human proteins targeted by viral proteins. The list of virus-targeted host proteins is especially long in the case of the HIV-1 virus, arguably the most extensively studied virus at present. The motif/counter-motif pairs involved in virus-host PPI are yet to be discovered in most cases [39, 40]. Discovery of motifs and hotspots on HIV protein sequences at the interface of virus-host PPI will allow us to discover previously unknown virus-host interactions. Knowledge of the interacting components in virus-host PPI will impact the development of drugs that modulate or break bonds between virus and host proteins [11, 12, 14] . Currently known linear motifs are not enough to unveil the grammar of the

crosstalk between HIV and host targeted proteins. A recent study by Evans and colleagues [40] has shown linear motifs from the ELM database [27] are poor predictors of interactions between HIV and human proteins and discovery of new motifs seems crucial.

1.6. Significance of the Study

This thesis mainly aims to identify the patterns present at the interface of a viral protein binding to a host protein. Results presented in this dissertation, will contribute a wealth of knowledge on HIV-1 virus-host protein interactions mediated by pairs of motifs and counter motifs. Computer aided rational design approaches identified a large number of low molecular weight inhibitors of protein-protein interactions and processes involving such interaction [12, 13]. One important key to success in the discovery of small molecules that block protein interactions has been the knowledge of the interface of interacting proteins. The structural coverage of protein complexes facilitates the discovery of small molecules for modulating protein-protein interactions [41]. The developments on drug design incorporating protein interactions could likely have an enormous impact on the treatment of viral infections.

Despite recent progress in antiretroviral combination therapies (HAART) against HIV-1, drug toxicity and the emergence of drug-resistant isolates during long-term treatment of HIV-infected patients necessitate a new look and alternative approaches. It has been shown recently that a set of position-specific motifs on the sequence of HIV-1 reverse

transcriptase is strongly correlated with poor response to antiretroviral therapy [42]. In other words, the presence and absence of such motifs at specific regions of the HIV sequence is highly predictive of response to therapy. A better understanding of virus-host proteome crosstalk may lead to the discovery of genome-wide variations in the host and the virus responsible for poor response to HAART therapies. Results presented in this thesis will provide insights on the potential use of new or existing drugs to treat neurological, cardiovascular and other ailments frequently observed in HIV-1 positive patients. One example along these lines is the use of cholesterol regulating drugs in the treatment of HIV-1. HIV-1 infection is known to be associated with altered lipid and lipoprotein metabolism and an increased risk of coronary artery disease. Bukrinsky and colleagues have recently shown that HIV-1 impairs ATP-binding cassette transporter A1 (ABCA1)-dependent cholesterol efflux from human macrophages. This effect was shown to be mediated by Nef [43]. Therefore, discovery of motifs involved in the binding of HIV-1 Nef to ABCA1 could be the key for developing an effective therapy blocking impairment of cholesterol flux.

The findings of the proposed research will have broad implications on viral infections of the human population. The motif/feature annotations developed in this study could potentially be applied to other virus-host interactions, such as hepatitis B and C. An important outcome of the proposed research will be its impact on computational predictions of virus-host protein interactions for viruses with available multiple sequence alignments. It should also be noted that the Simian Immunodeficiency Virus

(SIV) shares motifs with HIV-1 [44]. Findings of this thesis will have an impact in assessing the clinical value of experiments that utilize SIV as a model of HIV-1 such as those on the role of SIV Vpr on G2/M cell cycle arrest [44] as well as vaccine studies involving SIV [45, 46].

1.7. Thesis Organization

In chapter two, sequence hotspots of HIV proteins that are associated with the crosstalk with host hub proteins will be discussed. First I indentified the most important human protein (in terms of number of interaction partners) targeted by HIV. Then I searched for common patterns present on both HIV proteins and the neighbors of human proteins targeted by the corresponding HIV proteins. These motif patterns highlight hotspots on the HIV-1 proteins sequences that are crucial in the cross talkwith human proteins.

In Chapter Three, I focused on Nef which is a regulatory HIV-1 protein and plays a major role in altering signaling pathways of cells. I used the same approach as in Chapter Two to identify hotspots on Nef associated with interactions with all human proteins targeted by Nef. I also studied the potential co-operation of the hotspots (motifs) by looking at their co-occurrence among host proteins outcompeted by Nef.

Chapter Four, using iron dependent host cell mechanisms, illustrates how currently available network building techniques enable one to integrate patchy research literature into a portrayal of species crosstalk affecting modes of host protein networks.

Finally, Chapter Five concludes the thesis by summarizing the main aspects of the research presented in this dissertation. Next, future work that stems from this dissertation is discussed. I will highlight major significant achievements of this thesis and potential use of its results.

Chapter 2: HIV-1 sequence hotspots for crosstalk with host hub proteins

2.1. Background

Hub proteins in the human protein network undergo transient binding interactions with hundreds of interaction partners, as quantified in the Human Protein Reference Database (HPRD) [20]. Using protein-protein interaction data involving pathogen strains, Dyer et al. [47] illustrated the tendency of pathogen proteins to preferentially interact with host hub proteins. Recent bioinformatics studies also demonstrated a significantly greater propensity for HIV to interact with highly connected host proteins [38, 48]. Multiple and repeated domains were shown to be enriched in date hub proteins along with long disordered regions [49], suggesting a mechanism for their ability to undergo transient interactions. Pairs of strings of domains are highly predictive of hub protein binding to other host proteins in phosphorylation events [50], however, domain-motif interactions appear to dominate phosphorylation of HIV proteins by host kinases [51].

The HIV-1, Human Protein Interaction Database [52] identifies nineteen host hub proteins with at least one hundred neighbors as directly binding to one or more HIV-1 proteins. Some of these hub proteins phosphorylate their partners, while others cleave or recognize HIV protein sequences for nuclear localization. The high copy number of viral proteins in infected cells may lead to the out-competition of host proteins for their interaction with hub proteins as part of the topology of signaling and metabolic protein

networks [40]. To quantify the changes imposed on the host protein network by HIV, it would be important to identify the hotspots on HIV-1 protein sequences that are used to interact with hub proteins. Such hot spots could represent potential antiretroviral drug targets [12-14]. Moreover, sequence patterns of such spots could be used to identify host proteins outcompeted by viral proteins, which is in line with the concept of motif sharing for hijacking a host protein function [38, 39]. Viral proteins can mimic native interfaces and thus interfere with binding events in host protein networks [31].

In this study, the identity of HIV targeted host hub proteins were used as the input, along with sequences of their binding partners and the multiple alignments of HIV proteins, in order to identify hotspots along the viral protein sequences for binding to host hubs. Motivation for this aim comes from recent system-wide studies highlighting the importance of HIV targeted host hub proteins in the course of infection [11, 53, 54]. The formula used in the present analysis for identifying sequence hotspots is based on motif discovery and motif enrichment. It is well established that linear sequence motifs, 3 to 10 amino acids long, play important roles in transient binding interactions among proteins [8, 55]. However, eukaryotic linear motifs documented in the literature appear to be too general and ubiquitous to be discriminating between false positives and false negatives [40, 48, 51] .

The high throughput approach to motif discovery presented in this chapter is specific to HIV and host proteins. The goal is discovery of short linear protein motifs that are

highly statistically enriched among neighbors of host hub proteins and are highly conserved in the varying sequences of HIV proteins. If a motif is not highly conserved on known sequences of at least one HIV protein, it is likely that the motif is not essential to viral infectivity. Secondly, an HIV motif involved in binding to a hub protein is likely to be present on the sequences of host proteins competing with HIV for transient binding interactions with the hub protein. Indeed a previous study showed that this was the case for eukaryotic linear motifs [40].

Multiple methods and approaches have been developed for *de novo* motif discovery using protein sets and protein interactome datasets [29-34]. Discovery of correlated motifs on binding partners in an interactome subset reduces the discovery of motifs with no apparent function [34], but is not readily suitable to the present case of identifying motifs on large numbers of proteins interacting with the same hub. As in the correlated motif discovery approach, the approach used in this study utilizes protein-protein interactions, but the dataset that has been used in this study for motif discovery is highly asymmetric containing only nineteen hub proteins on one side and their more than a thousand binding partners on the other side.

SLiMFinder [30] was employed for *de novo* motif discovery in this context, as it is comprehensive, customizable and has extensive documentation. For each HIV targeted hub protein, the set of host proteins that interact with the hub protein were identified using HPRD and sequences of HIV proteins known to bind to the hub protein were

added to this list multiple. Such sequence sets containing hundreds of protein sequences were created for motif discovery. The resulting lists of motifs were further tested for their statistically enriched presence among hub neighbors in comparison to the HPRD proteins. Motifs that passed the test were further considered for their conserved expressions on hundreds of multiple alignments of HIV proteins known to interact with hub proteins. The approach presented in this chapter, identified discrete sets of hotspots on HIV protein sequences potentially involved in HIV - host hub interactions. Eukaryotic linear motifs that were conserved on HIV proteins and were highly enriched among the binding partners of hub proteins intersected with some of these hotspots. An extensive literature search of directed mutagenesis events showed functional validity of about a dozen hotspots with previously unknown motifs, indicating the biological context of the motif discovery presented in this study.

2.2. Methods

2.2.1. Data Acquisition

Human protein interaction data were downloaded from HPRD [20], Release 8, and HIV, human protein interaction data were obtained from [52] (accessed December 2009). Eukaryotic linear motif (ELM) patterns were collected from the ELM resource [27]. The HIV-1 Sequence Database (http://www.HIV-1.lanl.gov/) for subtypes A, B, C, and D (2008 version) was used to download multiple protein alignments of HIV proteins (Env, Gag, Nef, Pol, Rev, Tat, Vif, Vpr and Vpu).

2.2.2. Dataset preparation and motif discovery

Among the human proteins annotated as directly interacting with at least one HIV protein in the HIV-1, Human Protein Interaction Database, nineteen had at least 100 immediate neighbors in the HPRD database. The choice of 100 as a lower bound for the number of neighbors of a hub protein is arbitrary to some extent, as some known human hub proteins such as CDK1 have a lower number of binding partners. Preliminary studies showed that the automated approach used in this chapter for motif discovery required significant computing time with increasing numbers of sequence batches and increasing numbers of sequences and lengths of sequences in each batch. The choice was also guided by preliminary computations indicating that no new hotspots were annotated on the HIV sequence as the number of hub proteins considered reached nineteen. Interaction of these host proteins with HIV proteins was described in the HIV-1 human protein interaction database, either as binding or using words such as "phosphorylates" or "cleaves".

In motif discovery, I sought motifs that are conserved on multiple alignments of HIV proteins and are over-represented among proteins that share a common function, i.e., interacting with the same hub protein. Thus, the sequence set for motif discovery associated with a specified hub protein and an HIV protein consisted of the sequences of all host proteins binding to the hub protein, as well as sequences of the HIV protein equal in number to ten percent of the number of hub neighbor sequences. The HIV

protein sequences used in motif discovery were chosen randomly from the collection of sequences. For instance, the hub protein TP53 has 266 neighbors and it is known to bind to Nef. Therefore, 27 randomly chosen Nef sequences were added to the dataset. Repeated random selection of HIV-1 sequences in this manner did not result in any new motif discovery. In total, 42 datasets were created for different hub human protein interactions with HIV proteins.

The sequence datasets thus prepared were fed into the motif finding tool, SLiMFinder, for discovery of motifs ranging from 3 to 10 amino acids in length. The Blast e-value used in this tool was set to 1e-28. Other parameters for motif discovery in SLiMFinder were set to the default values in the tool manual. Motifs computed as output were first matched to human proteins to eliminate abundant motifs. Motifs present in more than one third of HPRD proteins were filtered. A previous study based on eukaryotic liner motif annotation showed that motifs that were ubiquitously present were poor predictors of HIV- host interactions [40].

2.2.3. Statistical enrichment

Statistical enrichment of discovered motifs among immediate neighbors of hub proteins was calculated by using the hypergeometric test against their background expression in HPRD. Any protein containing at least one copy of a motif was deemed as motif expressing. P-value cutoff of 0.005 was chosen to eliminate non-significant motifs. Another requirement for further annotation of the discovered motifs is their conserved

presence on the HIV-1 sequences. Motifs that were not present on at least 80% of all of the major subtypes of the corresponding HIV-1 protein sequence were removed. Since the approach used in this chapter is based on over representation of a motif among neighbors of a hub protein, only those motifs that were present on at least 20 percent of the neighbors of the hub protein under consideration were kept. Therefore, the final list of motifs for each hub-HIV-1 protein dataset contained motifs that are over represented and enriched among the neighbors of the hub protein, not abundant in the human proteome, and present on a vast majority of the sequences of HIV-1 proteins interacting with hub proteins.

2.2.4. Hotspots annotation with literature on directed mutagenesis

Discovered motifs that passed the processing described above were projected onto HIV-1 protein sequences, and many of them contained the same sequence segments. Such amino acid sequences comprised a list of hotspots. The intensity of the hotspot was deemed proportional to the number of hub proteins with motifs intersecting with the hotspot, normalized with respect to the number of hubs known to interact with the HIV-1 protein under consideration. Next, PUBMED abstracts were searched for directed mutagenesis studies involving mutations falling within the range of the identified motifs and hotspots. I also identified eukaryotic linear motifs that were conserved on HIV-1 and statistically enriched among the neighbors of the hub proteins with the same cutoffs

used in the motif discovery. These data were used to provide a biological context to the HIV-1 sequence hotspots for hub protein binding in this study.

2.3. Results

This study aims to discover linear protein sequence motifs shared by HIV protein sequences and a large subset of the immediate neighbors of host hub proteins targeted by HIV. Randomly chosen viral protein sequences were combined with the sequences of proteins known to interact with HIV-1 targeted hub proteins, one hub protein at a time. A motif discovery algorithm was used to identify motifs that are conserved on HIV-1 sequences and statistically enriched among neighbors of HIV-1 targeted hub proteins. Table 2 lists the gene IDs and gene symbols of these hub proteins, along with the number of immediate neighbors and the GO molecular functions of these neighbors. Also shown in the table are the identities of HIV proteins interacting with these hub proteins. HIV-1 Tat and Nef interact with 9 and Gag with 7 of the hub proteins listed in Table 2. HIV-1 targeted hub proteins considered in this study are most frequently either kinases or transcription factors.

Table 2. List of host hub proteins targeted by HIV

The table lists HIV targeted human proteins with more than 100 immediate neighbors in HPRD. Also listed are the numbers of neighbors of hub proteins and *GO Molecular Functions* enriched among neighbors compared to the set of HPRD proteins. The last column identifies the HIV proteins targeting the hub proteins.

Entrez ID	Symbol	Neighbors Count	GO Molecular Function	HIV-1 Protein Interactor
7157	TP53	266	TF, RNA binding, DNA binding	Nef, Tat
2033	EP300	210	TF activator	Tat, Vpr
6714	SRC	208	kinase, RNA binding	Nef
1387	CREBBP	198	TF activator	Tat, Vpr
5578	PRKCA	173	kinase, RNA binding	Gag, Nef, Pol, Rev, Tat
1457	SNK2A1	169	kinase, RNA binding	Gag, Pol, Rev, Vpu
5594	MAPK1	160	kinase, kinase binding, RNA binding	Gag, Nef, Rev, Tat, Vif
2534	FYN	154	kinase, RNA binding	Nef
5566	PRKACA	145	kinase, kinase binding, RNA binding	Gag, Nef
5295	PIK3R1	128	protein phosphatase binding	Nef
983	CDC2	119	kinase, RNA binding	Rev
5595	MAPK3	116	kinase, RNA binding	Tat, Vif, Gag, Rev
3725	JUN	116	TF, DNA binding	Tat
801	CALM1	114	phosphorylase kinase	ENV, Gag, Nef, Pol
7431	VIM	112	kinase binding	Pol
5970	RELA	111	kinase binding, TF	Tat
3932	LCK	105	kinase, kinase binding, RNA binding	Nef
5580	PRKCD	102	kinase, RNA binding	Pol, Tat
60	ACTB	101	kinase binding, RNA binding	Gag, Pol

2.3.1. HIV sequence hotspots for interaction with host hubs

Shown in Figure 1 is a typical result of motif discovery, presented for the sets of motifs potentially used for crosstalk with the top 3 hub proteins and their position on HIV proteins. This radar plot illustrates motifs on a1) Vpr in interaction with EP300; a2) Tat in crosstalk with EP300; b1) Nef interacting with TP53; b2) Tat interacting with TP53; and c) Nef binding to SRC. The radar plot shown in the figure organizes discovered

motifs on circles with a radius equal to the sequence distance from the start of the protein sequence to the start of the discovered motif. The figure shows that many of the discovered motifs that are rich in proline and related to the LIG_SH3 ELM pattern are spatially clustered along the sequence of the HIV protein Nef. I refrained from consolidating these motifs into one pattern, as the motifs shown may have slightly different functions, as illustrated by the multiple ELMs known to interact with SH1, SH2, and SH3 protein domains. Radar plots for other hub protein-HIV protein pairs indicated the presence of hotspots where multiple discovered motifs merged.

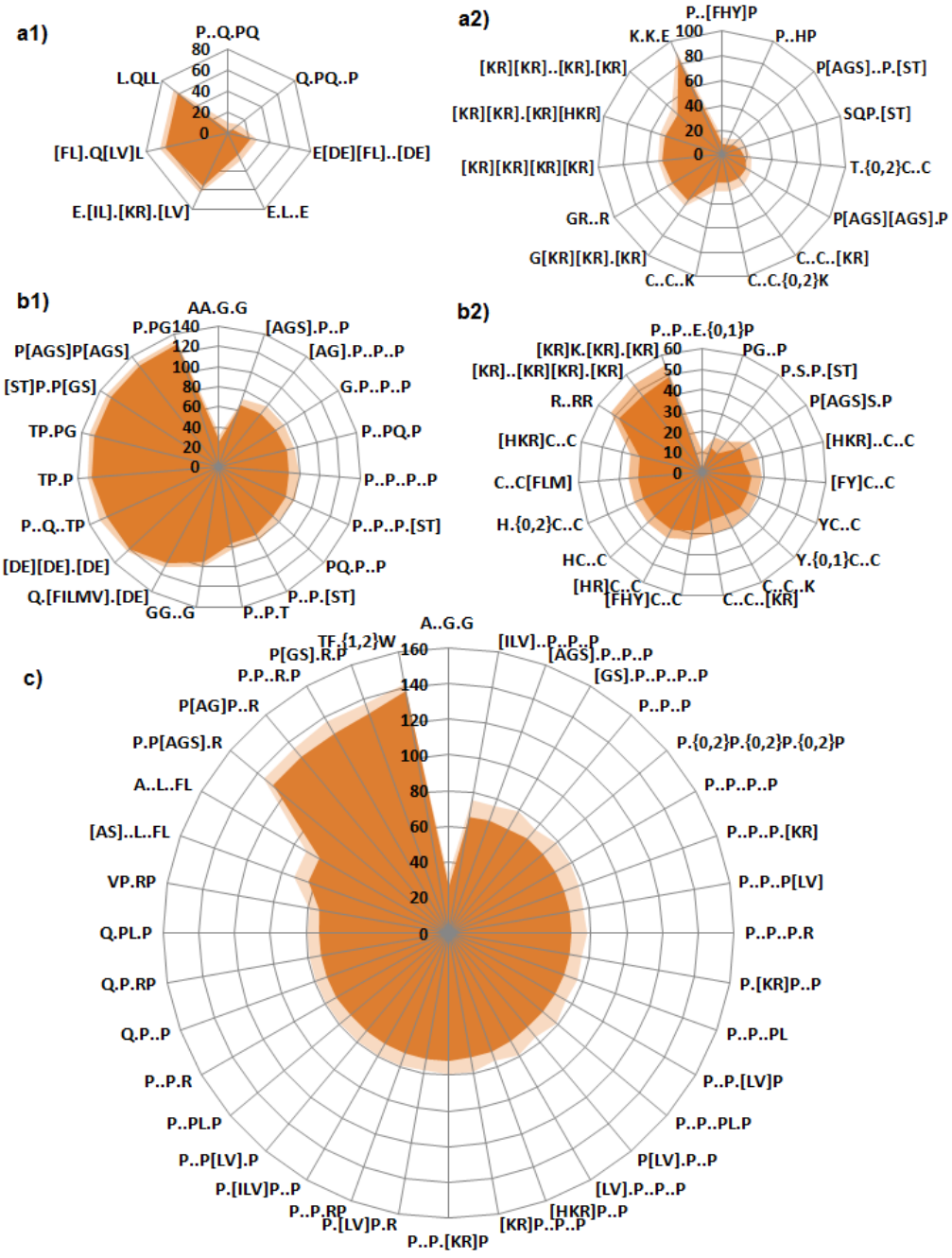

Figure 1. Motifs for top 3 hub proteins and their position on HIV proteins

This radar plot illustrates the HIV protein motifs discovered in the present study. a1) Vpr motifs for crosstalk with EP300 a2) Tat motifs for crosstalk with EP300, b1) Nef motifs for TP53, b2) Tat motifs for P53; and c) Nef for SRC motifs. More Detail on each motif is available in additional files. Sequence positions are mapped to distance from the radius. The dark edge represents the start of the motif and the lighter edge represents the end of the motif.

Next, the motifs that we discovered on the neighbors of multiple hub proteins were plotted on the HIV protein sequence, and the amino acids along the HIV protein sequence were marked with grayscale intensities proportional to the number of hubs associated with a motif on the hotspot. The resulting sequence hotspots for HIV proteins Tat, Rev, Nef, Gag, and Pol are shown in Figure 2. For achieving simplicity in the figure, only those hotspots with motifs from at least two hubs have been shown along the HIV proteins (horizontal axis) for hundreds of sequences ranging from 637 for Tat to 1792 for Gag. The figure shows increasing entropy on hotspot positions with increasing sequence length and sequence copy number. Aligning sequences for optimizing positional conservation required too many gap insertions and thereby distorted the actual positions of these motifs on the majority of the sequences for a given HIV-1 protein and thus this route was avoided. The figure shows four hotspots on Tat, five on Rev, eight on Nef, and significantly more on Gag and Pol.

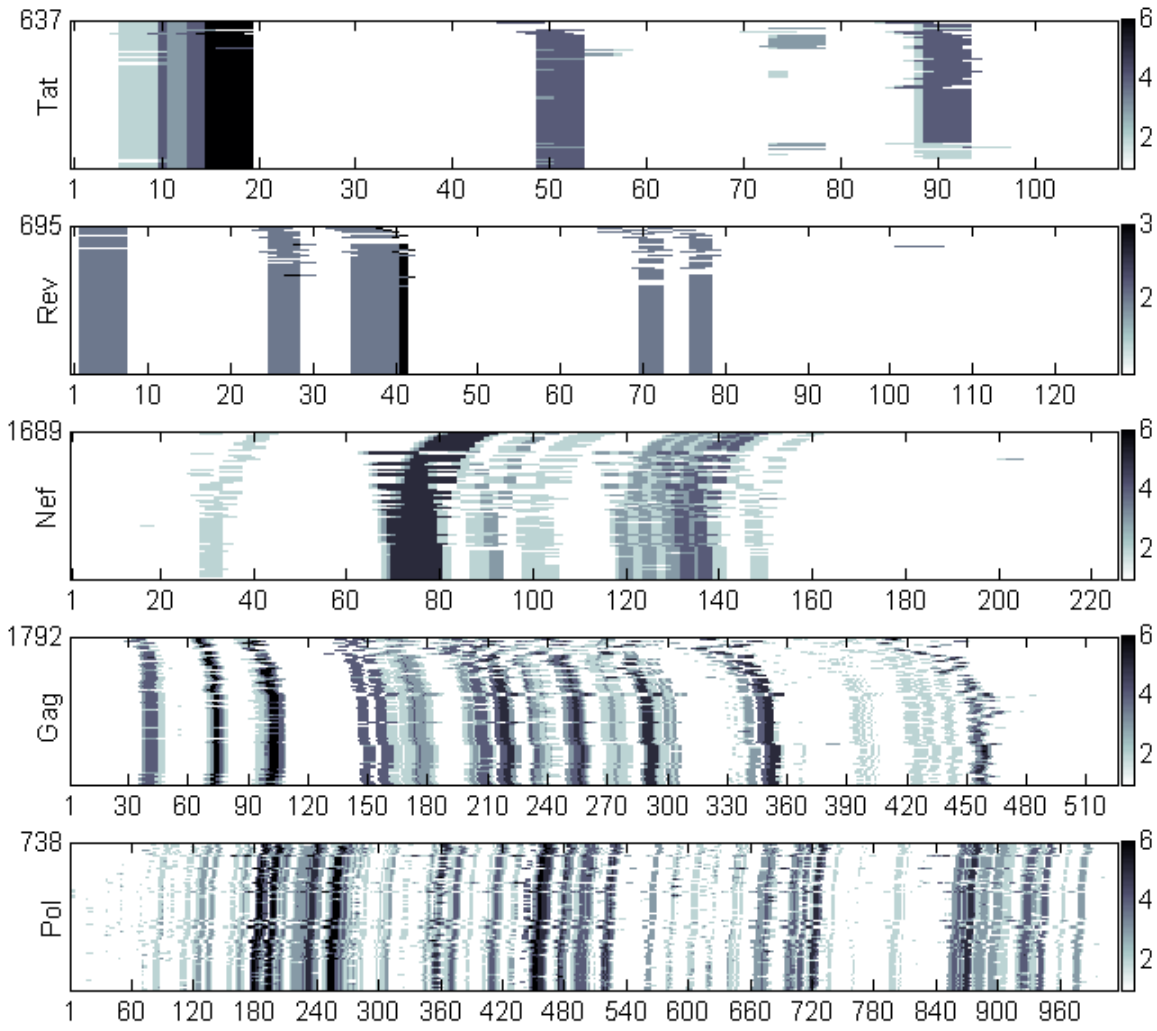

Figure 2. Motif hotspot positions on HIV protein sequences

Amino acid sequence positions of motif hotspots are shown on the horizontal axis. Color intensity is proportional to the number of hub proteins with enriched hotspot motifs among its immediate neighbors. Regions highlighted in this figure have at least two different hub proteins associated with them.

Next, I considered whether the hotspots shown in Figure 2 were mainly due to host hub proteins having large numbers of commonly shared neighbors. The heat map indicating the number of common neighbors for pairs of HIV targeted hubs is shown in Figure 3. Hotspots containing motifs from at least three hub proteins do not have enough

common neighbors to overcome the minimum 20% presence limit among hub neighbors imposed on motifs discussed in the present study. Similarly, hotspots composed of motifs from two hubs share few common neighbors, except in cases between MAPK1 and MAPK3, and also between EP300 and CREBBP. The hotspot shown in between position 2 to 7 on Rev is indeed due to MAPK1 and MAPK3 having common neighbors. However, this is a rare event among the hotspots shown in Figure 2 or the motif sets presented in Appendix 1.

	CSNK2A1	JUN	SRC	EP300	PRKCA	RELA	CALM1	PRKACA	TP53	CDC2	VIM	CREBBP	MAPK3	FYN	ACTB	LCK	PIK3R1	PRKCD	MAPK1
CSNK2A1	169	20	11	21	18	21	9	18	27	19	4	25	13	10	1	5	5	7	17
JUN	20	116	8	34	2	24	7	5	31	9	3	40	13	3	3	5	4	8	20
SRC	11	8	208	16	31	15	12	11	13	8	8	20	23	60	3	41	47	26	27
EP300	21	34	16	210	8	35	8	8	38	15	4	94	19	3	3	6	5	10	28
PRKCA	18	2	31	8	173	5	22	32	16	9	5	6	18	16	7	15	16	35	18
RELA	21	24	15	35	5	111	7	5	30	7	2	33	9	4	0	7	4	7	13
CALM1	9	7	12	8	22	7	114	16	5	5	1	10	5	9	2	6	3	9	8
PRKACA	18	5	11	8	32	5	16	145	7	7	5	9	14	6	2	6	8	12	19
TP53	27	31	13	38	16	30	5	7	266	24	11	32	12	6	5	10	9	12	16
CDC2	19	9	8	15	9	7	5	7	24	119	4	15	7	4	4	2	5	6	8
VIM	4	3	8	4	5	2	1	5	11	4	112	4	4	5	2	2	3	3	4
CREBBP	25	40	20	94	6	33	10	9	32	15	4	198	21	7	3	7	5	8	26
MAPK3	13	13	23	19	18	9	5	14	12	7	4	21	116	13	3	12	12	15	80
FYN	10	3	60	3	16	4	9	6	6	4	5	7	13	154	6	61	37	15	21
ACTB	1	3	3	3	7	0	2	2	5	4	2	3	3	6	101	3	5	8	3
LCK	5	5	41	6	15	7	6	6	10	2	2	7	12	61	3	105	32	14	17
PIK3R1	5	4	47	5	16	4	3	8	9	5	3	5	12	37	5	32	128	15	15
PRKCD	7	8	26	10	35	7	9	12	12	6	3	8	15	15	8	14	15	102	18
MAPK1	17	20	27	28	18	13	8	19	16	8	4	26	80	21	3	17	15	18	160

Figure 3. Heat map for common s among hub proteins considered in the study

The number of common immediate neighbors between two hub proteins is shown in the form of a square matrix. Color intensity is proportional to the number of protein neighbors common to two hub proteins.

2.3.2. Biological context for sequence hotspots

An important subset of the motifs presented in Appendix 1 corresponds to motifs already annotated by the ELM web server. Shown in Table 3 are the ELMs that satisfy the three conditions imposed on motif discovery, namely these ELMs are conserved along the HIV protein sequence, expressed infrequently on HPRD proteins, and are highly statistically enriched among the neighbors of hub proteins. The start and end positions of ELMs on HIV protein sequences is also shown in the table. The ELM motif LIG_SH3, a kinase associated motif, is present on HIV proteins Env, Gag, and Nef, whereas the nuclear localization signal motif TRG_NLS_MonoCore_2 is found on Tat and Pol. The PCSK cleavage site is found to be conserved on Rev and Pol. The immunoreceptor tyrosine-based switch motif is found expressed by Env. The fact that the method presented in this chapter, reproduced all eukaryotic motifs satisfying all stringent criteria (as explained in previous sections of the chapter) for motif annotation provides support against an extensive presence of false negatives in the motif discovery approach of this study.

Table 3. Eukaryotic linear motifs (ELMs) present on HIV and enriched among neighbors of hub proteins

The *Start* and *End* points of ELMs are specified based on the most commonly observed start and end points on the HIV protein sequences. The *p-values* listed stand for the statistical enrichment of the ELM among neighbors of the hub protein using a hypergeometric test.

HIV	ELM Name	Start	End	Hub Entrez ID	Hub Gene Symbol	p-value
Env	MOD_TYR_ITSM	35	43	801	CALM1	9.80E-04
	LIG_SH3_4	117	125	801	CALM1	5.30E-04
Gag	LIG_SH3_2	288	294	1457	CSNK2A1	2.79E-03
				5578	PRKCA	1.73E-03
				5594	MAPK1	1.92E-04
				5595	MAPK3	4.33E-04
	LIG_SH3_1	451	458	5594	MAPK1	4.24E-04
Nef	LIG_SH3_2	72	78	2534	FYN	2.67E-06
				5295	PIK3R1	6.08E-06
				5578	PRKCA	1.73E-03
				5594	MAPK1	1.92E-04
				6714	SRC	2.18E-11
Pol	CLV_PCSK_PC7_1	233	240	801	CALM1	7.79E-04
	TRG_NLS_MonoCore_2	255	261	1457	CSNK2A1	2.99E-03
Rev	CLV_PCSK_FUR_1	39	44	1457	CSNK2A1	4.01E-03
Tat	TRG_NLS_MonoCore_2	48	54	2033	EP300	2.99E-05
				3725	JUN	1.28E-03
				5970	RELA	1.18E-03
				7157	TP53	5.78E-04
				1387	CREBBP	1.88E-05
Vif	LIG_SH3_1	158	165	5594	MAPK1	4.24E-04

Next, a literature search was conducted on directed mutagenesis of HIV sequences and focused on 24 research articles presenting HIV mutations that intersect with the motifs discovered in this study. Fourteen of these mutations came with functional changes in HIV-host interactions as detailed in Table 3. For example, the hotspot positioned at residues 15-19 of Tat contained mutation S16A that is known to prevent Tat phosphorylation. The hub protein interacting with Tat at this position is PRKCD, a

kinase known to phosphorylate Tat. The second set of mutations (R52Q, R53Q) fell onto the hotspot intersecting with TRG_NLS_MonoCore ELMs, a motif recognized by the importer protein importin-alpha. The sequence hotspots on HIV proteins Vif, Vpr, and Vpu are not shown in Figure 3 for brevity, since there is no hotspot with more than one associated hub protein; however some of the discovered motifs on these proteins (presented in Appendix 1) intersect with mutations known to affect viral protein activity (Table 4).

HIV protein regions binding to some of the hubs shown in Table 2 were previously identified in the literature. Shown in Table 5 are twelve such regions that intersect with the hotspots shown in Figure 2 and another five that do not match with the hotspots. The approach of this study, discovered hotspots at the binding sites of Tat with CREBBP and EP300, but missed the Gag binding site to ACTB, and Rev to CDC2. One of the mismatches was due to the stringent criteria set for defining hotspots. The site of Vpu binding to CSNK2A1 was correctly recovered using the motif with regular expression [AGS]..S..E.[DE] in the sequence position 49 to 58 (Appendix 1) but this site was not shown in Figure 2 because it involved just one hub protein instead of two or more designations that was used for the definition of a hotspot.

Table 4. HIV amino acid mutations found in research literature within the range of motifs annotated in this study

The result of the literature search for the studied mutations on HIV-1 proteins that occur in the region of the discovered motifs is presented in this table. The related paper is listed in the *Pubmed ID* column. The mutation under consideration is presented in the Mutation column. Motif Pattern is the regular expression of the motif to which the mutation corresponds. Start and End of the motifs are specified based on the most commonly observed start and end on the HIV protein sequences.

HIV-1	Pubmed ID	Mutation	Motif Pattern	Hub Symbol	Start	End	Phenotype
Gag	9420228 [56]	P222A	A.{0,1}G.{0,2}P.{1,2}P	CSNK2A1	217	223	Diminishes virion incorporation of CyPA and interfere with HIV-1 replication
			P..PG	MAPK3	219	224	
Nef	10547288 [57]	F90R	[DE]L..[FIL][IL]	MAPK1	87	93	Reduces the affinity of SH3 binding (specifically in HCK, similarly in FYN, LCK)
			[DE]L..[FL]L	LCK	87	93	
			F.{0,2}L.{0,2}K	FYN	91	94	
	10489340 [58]	R77A	P..P.R	PIK3R1, SRC, MAPK1	73	79	Decreases downregulation of class I MHC
Pol	20450778 [59]	K101P	K.K.I	PRKCD	98	103	Correlated with drug resistance (failing combinational antiretroviral therapy)
		K219	A..KK	ACTB	215	220	
		K70R	[KR][ILV]..Q.[KR]	CALM1	68	75	
Tat	8709193 [60]	R52Q, R53Q	KR..R	RELA	51	56	Repeals Tat binding to TAR element and gene trasactivation
			[AS]..R.[KR][KR]	JUN	46	53	
			[KR][KR][KR][KR]	EP300, CREBBP	49	53	
	17083724 [61]	S16A	[FHY]..[ST].P	PRKCD	13	19	Prevents Tat phosphorylation and interferes with activation of HIV-1 provirus
Vif	8626571 [62]	S144A	S.Q.L	MAPK3	144	149	Loss of Vif activity (not phosphorylated)
Vpr	14506268 [63]	I61A,L64P	E.[IL].[KR].[LV]	EP300	57	64	Enhances pro-apoptotic activity of Vpr
	9557700 [64]	Q65E	QQL	CREBBP	65	68	Impairs Vpr nuclear localization
Vpu	20078884 [65]	S52A	[AGS]..S..E.[DE]	CSNK2A1	50	59	Interrupts phosphorylation of Vpu required for degradation of tetherin

Next, the hotspots were mapped to the 3D structures of three of the smaller HIV proteins under consideration. Tat, Rev, and Nef were retrieved from the protein data bank (PDB) [17], and hotspots on these structures were highlighted in orange in Figure 5. Note that structures in PDB are not complete and do not include the entire

sequences for Rev and Nef. Nevertheless, the figures clearly show that the identified hotspots do not form conformational recognition features. More likely, they are being utilized in anchoring two proteins at multiple sites. Motif combination sets used in binding events could be potentially quantified to some extent by statistical enrichment of their co-occurrence among the neighbors of the hub proteins targeted by HIV and not considered in this study.

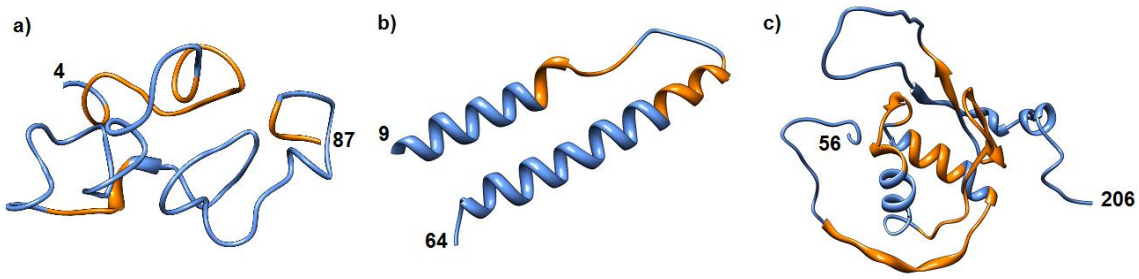

Figure 4. Hotspots on HIV protein structures

Hotspot regions highlighted in orange on Tat (a), Rev (b), and Nef (c) proteins. PDB structures 1TBC [66] , 2X7L [67], and 2NEF [68]were used respectively. Numbers on the structures reflect the start and stop positions on the actual HIV protein sequence. Molecular graphics images were produced using the UCSF Chimera package [69]

2.4. Discussion

The HIV alters the host cell macromolecule network and redirects cellular processes towards the synthesis of new viral particles. Binding interactions of HIV proteins with host proteins, DNA and RNA constitute a fundamental mechanism in the modification of host cellular networks in favor of viral production. Network connectivity is

significantly affected by the binding of viral proteins to host hub proteins. As shown in Table2, nineteen such host proteins with at least 100 binding partners appear as directly interacting with HIV proteins in the HIV-1 Human Protein Interaction Database. HIV-targeted host hub proteins are typically protein kinases and/or transcription factors. Therefore, alterations in their connectivity directly impacts signal flow through pathways and potentially leads to significant changes in global gene expression profiles.

Given that an HIV protein binds to a host hub protein, what can be said about the altered connectivity of the hub protein? One scenario would be that binding of the HIV protein to the hub protein occurs at sites utilized by host proteins to bind to the hub. Examples of such sites include phosphorylation and docking sites [51]. Even if phosphorylation of an HIV protein turns out to have little functional consequence on its own, the fact that multiple host proteins are outcompeted by the thousands of copies of the HIV protein would implicate a strong impact on network connectivity on the hub node under consideration. This is the rationale for the focus of the present study on the grammar of interactions between HIV and host hub proteins.

This study presents sets of newly annotated hotspots on HIV virus proteins as potential sites for binding to host hub proteins. The hotspots are at the intersection of short linear motifs shared by HIV proteins and the host proteins outcompeted by HIV proteins. A *de novo* motif discovery algorithm [30] was used with sequence data as the input, consisting of a hybrid of HIV-1 and host protein sequences, as described in methods. The output

consisted of motifs shared by the HIV and the host proteins competing in binding events to host hub proteins. As such, my computationally intensive motif discovery process used 42 sequence sets containing from a minimum of 111 to a maximum of 293 protein sequences. The motifs discovered in this study are (i) conserved on HIV protein sequences, (ii) found in less than one-third of the host proteins, and (iii) are statistically enriched among neighbors of host hubs targeted by HIV proteins. The sequence positions of these motifs on the HIV proteins constitute potential binding sites for host hubs. Thus, through a convoluted bioinformatics approach requiring extensive data on protein sequences and interactomes, the interface between HIV and host hub proteins was estimated.

The presented computational estimates of hotspots along the sequence of HIV proteins identified eukaryotic linear motifs associated with a nuclear localization signal on Tat and Pol, a PCSK mediated cleavage site on Rev and Pol, and a proline-rich kinase substrate motif on Env, Gag, and Nef (Table 3). In fact, this method reproduced all the eukaryotic linear motifs satisfying the stringent criteria imposed on their expression on HIV and on the neighbors of hub proteins. Findings of this study are also in line with large-scale experimental data on directed mutagenesis of the HIV sequences. Fourteen phenotype-altering single residue changes of HIV proteins collected from the literature were mapped onto the hotspot locations (Table 4). Additionally, the predictions of this study recaptured a large majority of the known interfaces between HIV and hub proteins (Table 5). To my knowledge, the large-scale motif analysis presented in this

study constitutes the first comprehensive map predictive of HIV-host hub binding interfaces. It was possible to create a hotspot map for the HIV proteome thanks to the extensive research findings in the literature on the identity of host hub proteins interacting with HIV proteins.

Potential uses of HIV sequence hotspots depicted in this study range from drug development to a better understanding of the mutation phenotypes in their linkage to host protein networks. Rational drug design procedures are increasingly focusing on developing drugs targeting protein-protein interaction interfaces [12]. The data produced by this study shows that the specific motif sequence segments expressed by viral proteins are often different than the motif sequences commonly used by the host. This provides an opportunity to block the binding interactions of HIV proteins and the host hubs using peptides or small molecules, without affecting hub connectivity to other host proteins. Another potential use is to provide a biological context for mutation phenotypes that maybe expressed in general terms, such as loss of viral infectivity [88].

Table 5. Available literature information on the binding sites of hubs and HIV-1 proteins binding interactions

Information available in the literature on the binding sites of the interaction between hub proteins and HIV-1 proteins is listed in the table. *Description* is the information available in the paper regarding the interaction site. *Pubmed ID* refers to the paper from which binding information was collected.

HIV-1	Hub Entrez ID	Hub Symbol	Description	Pubmed ID
ENV	801	CALM1	768 to 788 and 826 to 854 of gp41	8226798 [70]
GAG	60	ACTB	10 and 11	12009869 [71]
	5578	PRKCA	111	8473314 [72]
NEF	801	CALM1	1 to 20 (n-terminal)	15632291 [73]
	2534	FYN	65 to 82	7859737 [74]
	3932	LCK	69 to 78	8794306 [75]
	5295	PIK3R1	c-Terminal	12009866 [76]
	5566	PRKACA	6 to 9	15629779 [77]
	5578	PRKCA	nef 15	9049329 [78]
	5594	MAPK1	69 to 78	8794306 [75]
	6714	SRC	Proline-rich domain	16849330 [79]
	7157	TP53	1 to 57 (n-terminal)	11861836 [80]
REV	983	CDC2	14	8806671 [81]
	1457	CSNK2A1	5 to 8	
TAT	1387	CREBBP	1 to 24	12549909 [82]
	2033	EP300	48 to 57	11080476 [83]
	5578	PRKCA	46	8914829 [84]
	5580	PRKCD	46	
VIF	5594	MAPK1	96 and 165	9792705 [85]
	5595	MAPK3	96 and 165	
VPR	2033	EP300	64 to 84	12208951 [86]
VPU	1457	CSNK2A1	52 to 56	8548340 [87]

Hotspots link a phenotype altering mutation on an HIV protein to the identity of the host protein it interacts with at the site of mutation, allowing the use of bioinformatics in outlining a protein network pathway responsible for the phenotype. The motif collection presented in Appendix 1 is a comprehensive list of protein motifs shared by host hub neighbors potentially outcompeted by HIV. The size of the hub neighbor

protein set expressing a given motif provides a first order approximation of the identity of hub neighbors potentially outcompeted by HIV. A recently obtained crystal structure of HIV-1 Tat complexed with human P-TEFb provides further evidence that viral and host proteins interact on multiple sites, even in such rapid interaction events such as phosphorylation [89]. One could further refine the predicted outcompeted protein set by identifying those hub neighbor subsets enriched with an expression of multiple motifs positioned at different hotspots along the viral protein.

The motif sets presented in this study could be refined further by future bioinformatics studies utilizing structural information. Consideration of motifs within the context of a structural organization of proteins, such as their presence on helical loops [90] and disordered regions [91], may lead to a better understanding of the grammar of the HIV virus - host protein interactions and the role of short linear motifs in these interactions. Additionally, correlated motif approaches detailed in the literature [34] provide a map for identifying the interface on the hub protein interacting with a hotspot on the viral protein. Protein-protein interactions studied in this work for hotspot generation were asymmetric in the form of a single hub interacting with hundreds of proteins. The asymmetric sequence data used in this chapter were not suitable for a high throughput correlated motif approach. Now that HIV sequence hotspots are annotated with the motifs potentially used in binding to hub proteins, the recipe presented by Tan et al. [34] still provides a valuable opportunity to identify hub protein interfaces at hotspots. This study presents the first viral sequence hotspot map for the interaction of viral proteins

with the host hub proteins. The map is for the HIV proteome, the only viral proteome where extensive data exists concerning its communication with the host proteome. The hotspots on the viral protein sequences are potential binding sites to host hubs. Motifs that define the hotspots are indicators of the identities of host proteins outcompeted by viral proteins in their interactions with host hubs. The map presented in the study is highly consistent with experimental findings mined from the literature via text searching algorithms and subsequent manual curation. Findings of this study, impact well on drug development focusing on binding sites. Such findings will advance the knowledge regarding the details of the crosstalk between a virus and its host.

2.5. Conclusions

This study presents the first viral sequence hotspot map for the interaction of viral proteins with the host hub proteins. The map is for the HIV proteome, the only viral proteome where extensive data exists concerning its communication with the host proteome. The hotspots on the viral protein sequences are potential binding sites to host hubs. Motifs that define the hotspots are indicators of the identities of host proteins outcompeted by viral proteins in their interactions with host hubs. The map presented in the study is highly consistent with experimental findings mined from the literature via text searching algorithms and subsequent manual curation. The findings impact well on drug development focusing on binding sites. Such findings will advance our knowledge regarding the details of the crosstalk between a virus and its host.

Chapter 3: HIV Nef motifs for crosstalk with host proteins

3.1. Background

Viruses utilize host cellular mechanisms to redirect cell machinery toward viral replication. The process of viral replication includes multiple steps such as the incorporation of viral genome into host DNA, the synthesis of viral proteins using host machinery, and the assembly of viral particles. Viral proteins transiently and selectively bind to the DNA, RNA and the proteins of the host [92, 93] . The identification of host proteins targeted by viral proteins is important in outlining the progression of the infection and the key targets of the infection for developing virus- and patient-specific therapies and vaccine development.

Considering the fact that new viruses emerge in numbers and that advanced eukaryotes possess tens of thousands of proteins, identifying virus-host protein interactions in a wholesale manner is out of the question with the present experimental approaches. Such procedures combine multiple methods including co-immunoprecipitation, yeast two hybrid, protein arrays, phosphorylation assays, directed mutagenesis and others. Moreover, data generated from existing methods is noisy and the level of noise increases with generic, large-scale applications of such methods [94].

Recognizing the need for the discovery of the grammar used in virus-host interactions, a number of recent studies utilized machine learning algorithms, association methods,

and statistical enrichment to predict virus host protein interactions [40, 48, 51]. These studies annotated the viral sequences host motifs and domains previously associated with binding interactions in the host proteome. The short linear motifs are typically relatively disordered and flexible segments of proteins and as such may be involved in protein-protein interactions with more ordered molecular recognition features (MoRFs) on opposing proteins [22] . Molecular recognition features may be composed of amino acid residues not necessarily adjacent in sequence and could be parts of the closely packed regions of proteins called domains [95]. Because viral proteins are relatively flexible and contain long disordered segments, they may potentially express motifs targeting host proteins [38, 39].

A single motif-domain pairing may not be enough for inducing functional binding interactions; multiple bridges and interfaces are observed in 3D configurations of protein pairs [10]. Moreover, short linear motifs documented in open access bioinformatics databases such as the Eukaryotic Linear Motif (ELM) database [27] present regular expressions of over 140 motifs, expressed by eukaryotes. Prediction of PPI with currently annotated motif-domain pairings result in statistically significant intersections with experimental data [40, 48]. However, the presence of large numbers of false positives and false negatives in this approach creates a challenge in evaluating the directional changes in signal flow in host protein networks caused by the virus. It appears that data on viral sequence and host interactome may not be sufficient at the present to decipher the grammar of the crosstalk between viral and host proteins.

Recent advances in motif discovery on protein sets could be translated into the discovery of motifs specific to virus host interactions. Viral protein sequences of retroviruses such as HIV is highly variable but collection of such sequences alone is not sufficient for motif discovery since such an attempt typically yields thousands of motifs consisting of sequences with low entropy. In the previous chapter, a scheme was proposed for viral protein motif discovery using data on the binding interactions of virus host proteins. In this scenario viruses hijack the function of host proteins by outcompeting them in binding to other host proteins. It is assumed that viral proteins use motifs to interact with the interfaces on host proteins used for binding to other host proteins [14, 39, 44, 74]. Therefore motif discovery is carried out on protein sets composed of host proteins potentially outcompeted by viral proteins. Host protein set composed of all binding partners of host proteins targeted by the virus are enriched with outcompeted host proteins and as such comprise a set for motif discovery. Adding random samples of viral protein sequences assures that at least a significant subset of motifs discovered exist on the viral sequence as well as a subset of outcompeted proteins.

Protein features on the human proteome are better documented than viral proteomes in general. A number of open access web tools are available to annotate human proteins with motif regular expressions and protein domains [27, 32, 96, 97]. Moreover, recent bioinformatics studies have already associated pairs of protein signatures on opposing proteins to protein-protein interaction events [32]. In this study, the aforementioned

bioinformatics tools were used to make advances toward deciphering the grammar of HIV Nef-host protein interactions.

HIV Nef is a regulatory viral protein present in primate lentiviruses. It has a strong impact on the optimal replication of the virus, playing a major role in the transcription and translation of viral proteins. The Nef-protein is expressed abundantly in the early stages of viral replication [98, 99]. It possesses a structurally flexible N-terminal membrane anchor region of 60 residues, followed by a well-conserved and folded C-terminal core domain of about 130 residues. A flexible loop, 30 amino acids long, projects out of the core domain. Nef is post-translationally modified by phosphorylation and by the irreversible attachment of myristic acid to its N-terminus [98, 99] .

One of the main functions of Nef in viral replication is altering the signaling pathways of cells by interacting with tyrosine and serine/threonine kinases [77]. Nef increases the infectivity of the virus after entry into the cell [100]. Its interaction with the components of endocytic machinery decreases the expression of CD4 and major histocompatibility complex class I (MHC I) antigens on the surface of infected cells [101]. The molecular mechanism of Nef-mediated downregulation of CD4 and MHC I molecules are relatively well understood, but correlating these functions with the pathogenesis of AIDS is not straightforward [102]. Downregulation of CD4 may be important for optimal virus replication and it may facilitate the release of virions. Molecules that block the main interaction sites of Nef could be useful therapeutic agents [99] .

The HIV-1, Human Protein Interaction Database (HHPID) [52] identifies more than nineteen human proteins targeted by Nef. The immediate neighbors of the Nef targeted host proteins can be identified to an extent using existing human protein interactome databases such as HPRD [20]. Moreover, in much of these interactions, experimental data is available either on motifs involved or the sequence positions of the interface between Nef and host proteins, allowing for a critical assessment of the proposed system approach. The results show high consistency with experimental data and provide insights into motif combinations used by viral proteins in targeting host proteins.

3.2. Methods

3.2.1. Data Acquisition

Human protein interaction data were downloaded from the Human Protein Reference Database (HPRD) [1], Release 8. The HIV, Human Protein Interaction Database (HHPID) [8] (accessed December 2009) was used to obtain HIV Nef – host protein interactome. Host proteins targeted by HIV Nef were identified as those that are listed in HHPID as directly binding to or co-localized with Nef. In addition, the research literature linked to HHPID was screened to prepare summary tables of experimental data concerning motifs and sequence segments involved in host - Nef protein interaction events. These tables were then used to assess the overall consistency of the results with the research literature. The HIV-1 Sequence Database (http://www.HIV-1.lanl.gov/) for subtypes A,

B, C, and D (2008 version) was used to download multiple protein alignments of the HIV protein Nef.

Human protein sequences from NCBI's GenBank were screened for annotation of protein domains using PROSITE (version 20.31) [97]. Similarly, the Eukaryotic linear motif (ELM) resource was used to annotate protein sequences ELMs. The same set of sequences was screened for motifs discovered in the present study using regular expressions of the motifs.

3.2.2. Dataset preparation

In this study, 42 human proteins in HHPID were identified as participating in binding, phosphorylation, and cleavage interactions with HIV Nef. All three of the chosen interaction types involve direct binding of Nef to human proteins. Nineteen out of the forty-two had at least twenty direct partners (Table 6). These proteins were scanned against PROSITE [97] domain patterns and results are depicted in Figure 6. In total, 27 different domains are present on the list of 19 human proteins targeted by Nef. Among these domains, six are either signatures of kinases or Src homology profile which is in line with the fact that Nef is getting phosphorylated within the cell to alter major signaling pathways of the cell. No PROSITE domain was observed on AP2B1 and GNAO1.

Table 6. Human Proteins Targeted by Nef with interactions involving protein binding

entrez ID	gene Symbol	Neighbors	entrez ID	gene Symbol	Neighbors
7157	TP53	266	1315	COPB1	17
6714	SRC	208	162	AP1B1	17
5594	MAPK1	160	164	AP1G1	17
2534	FYN	154	8726	EED	17
5295	PIK3R1	128	3107	HLA-C	14
801	CALM1	114	3106	HLA-B	14
3932	LCK	105	5478	PPIA	13
5894	RAF1	93	3135	HLA-G	10
7409	VAV1	62	55690	PACS1	10
10399	GNB2L1	59	2625	GATA3	8
3055	HCK	56	8943	AP3D1	8
5062	PAK2	42	10318	TNIP1	8
375	ARF1	39	3133	HLA-E	7
4217	MAP3K5	38	8906	AP1G2	6
920	CD4	33	942	CD86	5
2775	GNAO1	33	3134	HLA-F	5
163	AP2B1	32	1174	AP1S1	5
3105	HLA-A	22	941	CD80	4
8907	AP1M1	21	8905	AP1S2	3
10015	PDCD6IP	19	51606	ATP6V1H	2
2623	GATA1	18	5692	PSMB4	2

In this chapter, the nineteen proteins (H1 proteins) were focused on for motif discovery as the approach presented in this chapter works best for sets of proteins with large amounts of binding partners. In motif discovery, each H1 protein was considered separately and sought motifs that are over-represented among its neighbors (H2 proteins). These H2 proteins share a common property which is interaction with the H1 protein. In addition to H2 proteins, Nef interacts with the H1 protein and the goal was to discover motifs common between Nef and H2 proteins. As a result, Nef sequences were

added to the dataset of H2 sequences. Nef sequences were selected randomly and their number was proportional (10%) to the number of H2 sequences in the dataset under consideration. Repeated random selection of Nef sequences in this manner did not result in any new motif discovery. In total, 19 datasets were created for the discovery of motifs potentially involved in the binding interactions of Nef and its host protein partners.

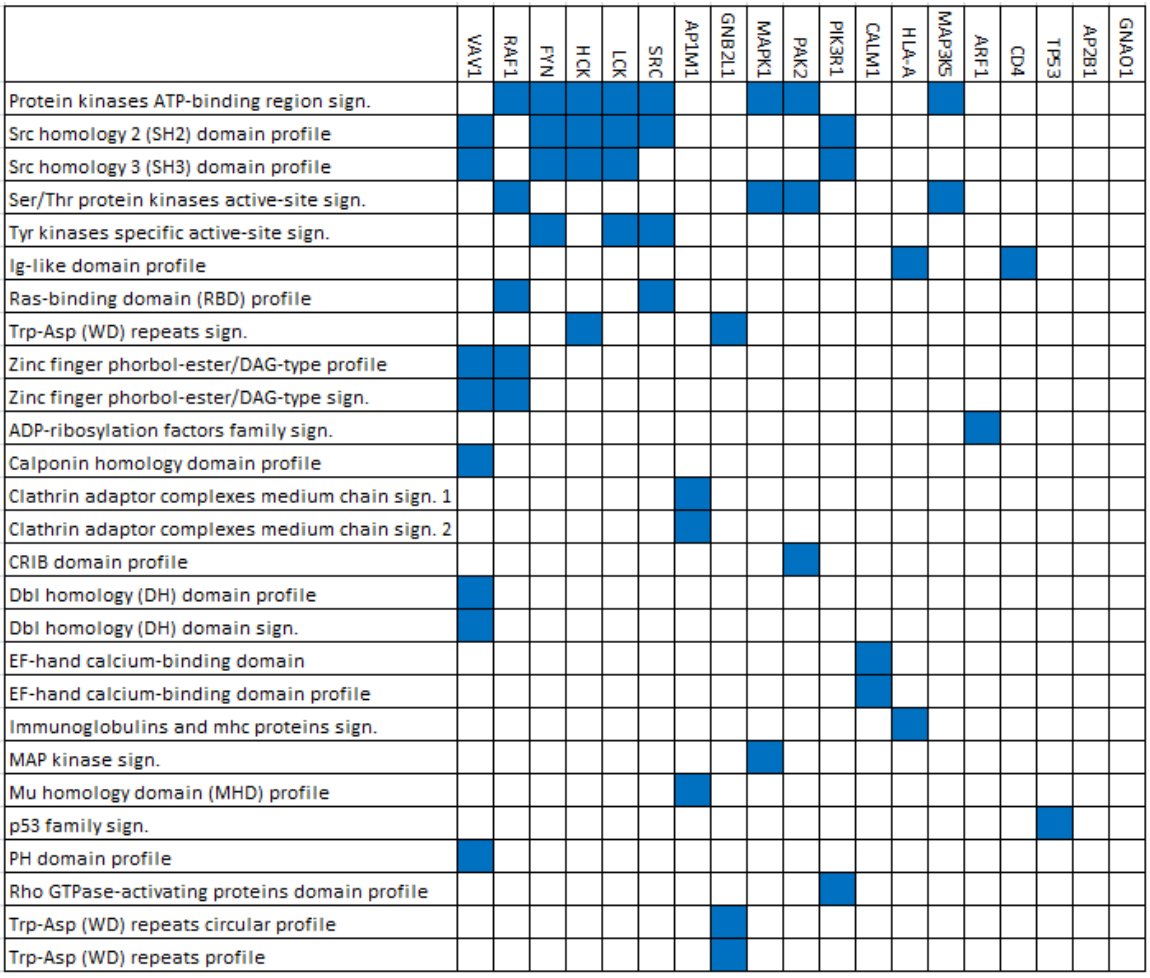

Figure 5. PROSITE domain annotations for top 19 human protein targeted by nef

The sequence datasets thus prepared were fed into the motif finding tool, SLiMFinder, for discovery of motifs ranging from 3 to 10 amino acids in length. The Blast e-value used in this tool was set to 1e-28. Other parameters for motif discovery in SLiMFinder were set to the default values in the tool manual. Motifs computed as output were first matched to human proteins to eliminate abundant motifs. Motifs present in more than one third of HPRD proteins were filtered. A previous study based on eukaryotic liner motif annotation showed that motifs that were ubiquitously present were poor predictors of HIV- host interactions [40] .

3.2.3. Statistical enrichment

The statistical enrichment (over-representation) of the discovered motifs was calculated among immediate neighbors of hub proteins by using the hypergeometric test against their background expression in HPRD. Any protein containing at least one copy of a motif was deemed as motif expressing. A p-value cutoff of 0.001 was used to eliminate potentially insignificant motifs. Another requirement for further annotation of the discovered motifs is their conserved presence on the HIV Nef sequences. Motifs that were not present on at least 80% of all of the subtypes of the Nef protein sequences were removed. Therefore, the final list of motifs contained motifs that are over represented among the neighbors of Nef targeted proteins, not abundant in the human proteome, and present on a vast majority of the sequences of HIV Nef proteins in the sequence database.

3.2.4. HIV Nef sequence hotspots for crosstalk with host proteins

The H2 motifs that passed the processing described above were projected onto HIV Nef protein sequences. Many of these motifs partially overlapped on the sequence. These motifs were clustered automatically based on their location on the Nef sequence. The developed clustering algorithm is an iterative merging algorithm. At the beginning each motif is considered as a cluster. Two clusters are merged if the sum of the mismatch between their beginning and ending positions (on Nef) is less than six amino acids. Merging was done iteratively through the motif list until no cluster changed. Each such cluster was labeled by the start and the end of the region on Nef covered by the cluster. These clusters represent hotspots on Nef. Each hotspot corresponded to a subset of H1 proteins for which a motif belonging to the hotspot was enriched among the immediate neighbors of these H1 proteins. Thus the subsets of H1 proteins potentially interacting with Nef via the hotspot were identified by this study.

Next, for each H1 protein interacting with a hotspot, the subsets of H2 proteins were identified by requiring the members of the subsets to express at least one motif clustered on the hotspot. The subsets of H2 proteins thus computed provided lists of host proteins potentially outcompeted by HIV Nef as a function of the hotspots expressed by Nef. Gene ontology molecular function level four was determined as a function of Nef hotspots for both and H2 proteins potentially outcompeted by Nef.

3.3. Results

The motif discovery approach used in this chapter computes Nef sequence motifs potentially targeting host proteins in an indirect fashion: I actually conduct motif discovery on the sequences of host proteins neighboring Nef targeted host proteins. The hypothesis is that Nef copycats motifs on host proteins used in their binding interactions. Since Nef is relatively flexible and disordered compared to host proteins in general, it is feasible that a disordered short linear segment of Nef binds to an already established interface position on a host protein. Small numbers of Nef sequences were added into the set of sequences used in the motif discovery to assure that some of the output motifs are expressed by the majority of Nef sequence collections used for verification. The host proteins targeted by HIV Nef were identified from HHPID and are shown in Table 7 along with the gene symbols/names for these proteins, and information the on number of host protein neighbors these proteins have. The table shows 19 of the Nef-targeted host proteins as having at least twenty immediate neighbors ranging from 266 neighbors for TP53, 208 for SRC and 160 for MAPK1 to 20 for AP1M1. For each of these nineteen proteins, the motif discovery tool was run and determined a set of motifs statistically enriched on H2 proteins of the given H1 as well as expressed by at least 80 percent of the multiple alignments of HIV Nef in the collection of Nef sequences. Because there is a significant range difference in the numbers of direct neighbors of Nef-targeted host proteins, an additional restriction was

added which is the motif discovered must be statistically enriched within the set of H2 proteins for all nineteen H1 proteins compared to host proteins in HPRD.

The results indicate multiple hotspots on Nef for interacting with the host proteins. Table 7 shows the HIV Nef short linear motifs that passed the aforementioned criteria. The table presents the regular expression (patterns) of such motifs along with their start and end points along the Nef sequence. Also shown in the table are the identities of H1 proteins for which the motif is statistically enriched on the set of its binding partners. Additionally, for each hotspot, statistically enriched GO molecular functions (level 4) of human proteins expressing any of the motifs in the cluster were computed and added to Table 7. Overall, there are ninety such patterns and some are enriched among the neighbors of multiple H1 proteins. These motifs were clustered according to their sequence position on Nef, resulting in 20 such clusters, with the largest one containing the proline-rich motifs spanning Nef residue positions 69 to 81. Next in abundance are the leucine-rich motifs in the sequence region from residue 85 to residue 96. The twenty clusters thus formed create hotspots on HIV Nef for potential binding interactions with host proteins (Figure 7).

Table 7. Motif Clusters (hotspots) on Nef

The table lists the motifs enriched among the immediate neighbors of human proteins targeted by Nef. These motifs are clustered based on their start and end on the Nef sequence. The *ELM* columns corresponds to the most similar ELM to the motif pattern. *GO_MF4*is a list of top GO molecular level 4 annotations for the human proteins containing the corresponding motif.

cluster	motif	start	end	associated h1s	ELM	GO_MF4
1_8	[IMV]..K.[GS][HK]	1	8	AP1M1	-	adenyl nucleotide binding, purine ribonucleotide binding, kinase
13_19	W.{1,2}A.{0,2}E	13	19	VAV1	-	phosphotransferase, kinase, adenyl nucleotide binding
26_35	[AS]..[GS]..[AGS].S	26	35	AP1M1	LIG_WH1	transcription factor , adenyl nucleotide binding, kinase
31_37	[GS][AS].[ST].[DE]	31	37	CD4	-	transcription factor , kinase , adenyl nucleotide binding
37_42	L.K.G	37	42	TP53	-	adenyl nucleotid binding, purine ribonucleotide binding, kinase
44_48	[ST][ST].N	44	48	FYN	-	purine ribonucleotide binding, transcription factor , kinase
66_72	[DE][LV][GS]F	66	70	PAK2	-	purine ribonucleotide binding, adenyl nucleotide binding, kinase
	[DE].[GS][FH].[LV]	66	72	AP1M1	-	

Table 7. (continued)

cluster	motif	start	end	associated h1s	ELM	GO_MF4
69_85	F.[FILV]..Q	69	75	PAK2	-	Kinase , phosphotransferase , purine ribonucleotide binding
	[FV]..[KR]P.[IV]	69	76	AP1M1	-	
	P..P..P[FILV]	70	78	HCK	LIG_SH3_1	
	P.{0,2}P.{0,2}P.{0,2}P	70	80	MAPK1, FYN, HCK, LCK, SRC, PIK3R1	LIG_SH3_1	
	V.P..P	71	77	ARF1	LIG_SH3_1	
	[HKR]P..P	72	77	FYN, VAV1, SRC	LIG_SH3_1	
	P.VP	73	77	FYN, HCK, VAV1	LIG_SH3_1	
	P..P.[HKR]	73	79	FYN, VAV1	LIG_SH3_2	
	P..P.R	73	79	MAPK1, FYN, HCK, VAV1, SRC, PIK3R1	LIG_SH3_2	
	Q.P.[HKR]	74	79	FYN	-	
	Q.P..P	74	80	PAK2, FYN, HCK, LCK, SRC	LIG_GYF	
	PL.P	75	79	ARF1	LIG_GYF	
	[FLV]P..P.T	75	82	HLA-A	LIG_SH3_1	
	P.[KR]P	76	80	VAV1	LIG_SH3_1	
	P.[HKR]P	76	80	VAV1	LIG_SH3_2	
	L.P.T	76	81	GNB2L1	-	
	P[LMV].P.[ST]	76	82	HCK	LIG_SH3_1	
	P..P.T	76	82	PAK2, FYN, HCK, HLA-A	LIG_SH3_1	
	P..P.[ST]..[AG]	76	85	HLA-A	LIG_SH3_1	
	P..P.[ST]..[AGS]	76	85	MAPK1, CALM1	LIG_SH3_1	
80_92	[LM][ST][FY]..[AG]..[FIL]	80	89	PAK2	-	guanyl-nucleotide exchange factor, SH3/SH2 adaptor, small GTPase regulator
	[HK]..[FLV].L..[FY]	83	92	ARF1	-	

Table 7. (continued)

cluster	motif	start	end	associated h1s	ELM	GO_MF4
85_96	[AS]..[FL]..[FLM]..E	85	95	MAP3K5	LIG_BRCT_BRCA1_1	purine ribonucleotide binding, kinase , phosphotransferase
	[DE]L[GS][FH]	87	91	PAK2	-	
	[DE]..[FHY][FY]	87	92	FYN	LIG_WH1	
	[DE]L..[FL]L	87	93	LCK	TRG_LysEnd_APsAcLL_1	
	D..[FH][FL][LV]	87	93	HCK	TRG_LysEnd_APsAcLL_1	
	[DE]L..[FIL][IL]	87	93	MAPK1	TRG_LysEnd_APsAcLL_1	
	[DE][ILM]..[FIL]..E	87	95	RAF1	-	
	L..FL	88	93	HLA-A	-	
	L..[FIL]L[KR]	88	94	VAV1	LIG_NRBOX	
	[FIL]..[FI][IL]..[KR]	88	96	HLA-A	-	
	[FIL]..F[IL]..[KR]	88	96	HLA-A	-	
	[IL]..[FY]L..[HKR]	88	96	ARF1	MOD_TYR_DYR	
	[GS].[FIL][LV]K	89	94	ARF1	-	
	[FH]FL..[KR]	90	96	HLA-A	-	
90_102	[FHY]..K.[KR]..[FLV]	90	99	MAP3K5	LIG_MAPK_1	adenyl ribonucleotide binding, protein kinase , pyrophosphatase
	F.{0,2}L.{0,2}K	91	94	FYN	-	
	F.KE	91	95	ARF1	-	
	FL..K	91	96	HLA-A, VAV1	-	
	[FWY]..E..G..G	91	101	PAK2	-	
	L..K.G	92	98	AP2B1	-	
	[GS].L[DE].[FL]	96	102	AP1M1	-	
108_113	Q.{1,2}I.{0,2}D	108	113	ARF1	-	adenyl nucleotide binding, purine ribonucleotide binding, kinase
	[DE][ILMV][LV][DE]	109	113	AP2B1	-	
	I.{0,2}L.{0,2}D	110	113	VAV1	-	

Table 7. (continued)

cluster	motif	start	end	associated h1s	ELM	GO_MF4
111_117	[LV].[IL][WY].Y	111	117	PAK2	-	phosphotransferase, kinsae , purine ribonucleotide binding
	[DE].[FWY].[FY]	112	117	RAF1	-	
118_126	Q.{0,2}F.{0,1}D	118	124	AP2B1	-	metal ion binding, adenyl nucleotide binding, purine ribonucleotide binding
	QG.[FILV]P	119	124	PAK2	-	
	Q..[FY].D	119	125	AP1M1	-	
	Q..[FWY]..[FWY]	119	126	CALM1	-	
	Q..[FHY]..[FWY]	119	126	CALM1	-	
	[GS][FY][FI]P	120	124	AP1M1	-	
	G[FWY][FIV]P	120	124	PAK2	-	
	G[FY][FI]P	120	124	PAK2	-	
121_132	[FY]..[DE][FWY]	121	126	MAP3K5	-	purine ribonucleotide binding, adenyl nucleotide binding, kinase
	[FHY]..[DE]..N	121	128	AP2B1	-	
	P..Q.[FY]	123	129	GNB2L1	-	
	[DE].Q.[FWY]	124	129	FYN	LIG_CAP-Gly_1	
	[DE]..N.[ST]	124	130	FYN, LCK	-	
	[FWY].N.[ST]	125	130	CD4	LIG_SH2_GRB2	
	W.{0,2}Y.{1,2}G	125	132	CD4	-	
	N[FHY][ST]	127	130	GNB2L1	-	
130_138	P.P..[HKR]	130	136	FYN, HCK	LIG_GYF	phosphotransferase, transcription factor, kinase
	P.P..[HKR].P	130	138	HCK	-	
	P..[HKR].P	132	138	MAPK1, FYN	-	

Table 7. (continued)

cluster	motif	start	end	associated h1s	ELM	GO_MF4
136_141	[FY]P..[FHY]	136	141	AP1M1, PAK2	-	protein domain specific binding, enzyme binding, purine nucleotide binding
	[FY]P..[FY]	136	141	AP1M1	-	
	[FWY]P..[FY]	136	141	LCK	-	
139_145	T.{0,2}F.{0,1}W	139	143	GNAO1	-	phosphotransferase , kinase , ion transmembrane transporter
	[FW][GS]..[FWY]	140	145	MAP3K5	-	
160_169	EN..L	160	165	TP53	-	peptide receptor, adenyl nucleotide binding, transmembrane receptor
	N..[IL].P	163	169	HLA-A	-	
200_206	[FHY]P..[FY]	200	205	PAK2	-	Kinase , phosphotransferase , purine ribonucleotide binding
	PE.{0,1}Y	201	204	HLA-A	-	
	[DE][FY][FWY]	202	205	AP2B1, RAF1	-	
	[DE][FWY][FY]	202	205	AP1M1	-	
	[DE][FY][FY]	202	205	AP2B1, RAF1	-	
	[FWY][FY][HK]	203	206	VAV1	LIG_WH1	

The majority but not all of the 1689 sequences in the multiple alignment of Nef express the motifs shown in Figure 7 within the twenty aforementioned hotspot positions. Shown in Figure 9 are the motif clusters on Nef along the 3D crystal structure. The start position of the hotspots is marked on the 3D structure. It is known that Nef core regions have multiple configurations depending on the myristylation of its N terminal region [103]. All the interactions mediated by the regions highlighted in the figure cannot happen simultaneously, because the binding sites are overlapping. It is probable, that different interactions occur at different stages of the viral life cycle. Nef may have a closed conformation, where the unsaturated parts are protected, and an open conformation, where the interaction sites are at hand.

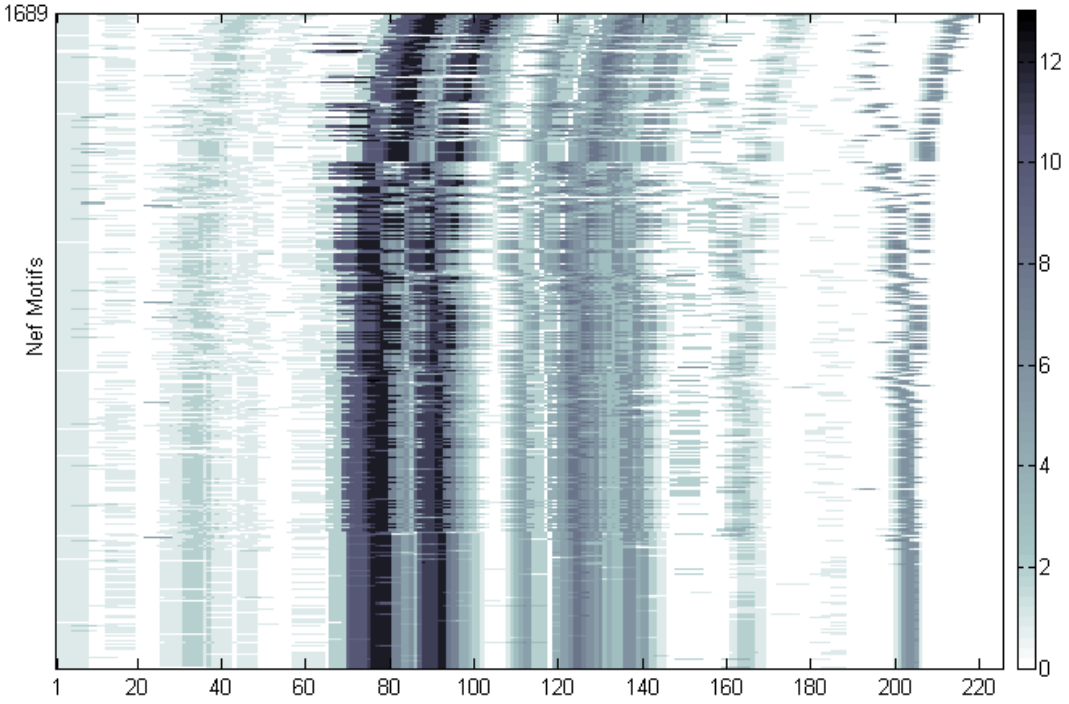

Figure 6. Motifs presence on Nef sequences

Amino acid sequence positions of motif hotspots are shown on the horizontal axis. Color intensity is proportional to the number of Nef targeted proteins with enriched hotspot motifs among their immediate neighbors.

The results indicate the presence of multiple binding sites on Nef for the host proteins it targets. Shown in Figure 8 are the hotspots linked to Nef targeted host proteins. The links in blue indicate computational predictions of this study with no experimental verification whereas those in purple indicate similar predictions with experimental support from directed mutagenesis and other methods as listed in HHPID. The figure shows that the developed bioinformatics approach correctly recaptures the functional role of the proline-rich Nef hotspot in binding to FYN, GNB2L1, HCK, MAPK1, PK3R1, SRC, and VAV1. Statistical method assessment of the match is not suitable in the current state of meager and noisy data on Nef – host protein binding interface sites. Nevertheless, the present findings bring a rationale for design of the high throughput experiments for elucidating information on HIV virus host protein interaction sites.

Next the lists of host proteins potentially outcompeted by Nef at the hotspots were determined and depicted in Figure 7. For this purpose, the immediate neighbors of the hotspot interacting H1 proteins expressing a motif falling onto the hotspot were identified. DAVID bioinformatics tools [104] was used to annotate the gene ontology molecular function assignments of these proteins as shown in Table 7. Some of the motifs in the twenty clusters on Nef hotspots have patterns that intersect the previously annotated eukaryotic linear motifs presented in the ELM web tool. A case in point is the proline-rich motif described by the regular expression P..P. Also shown in Table 7 are the ELM motifs deemed similar to the motifs on the hotspots using the CompariMotif [105].

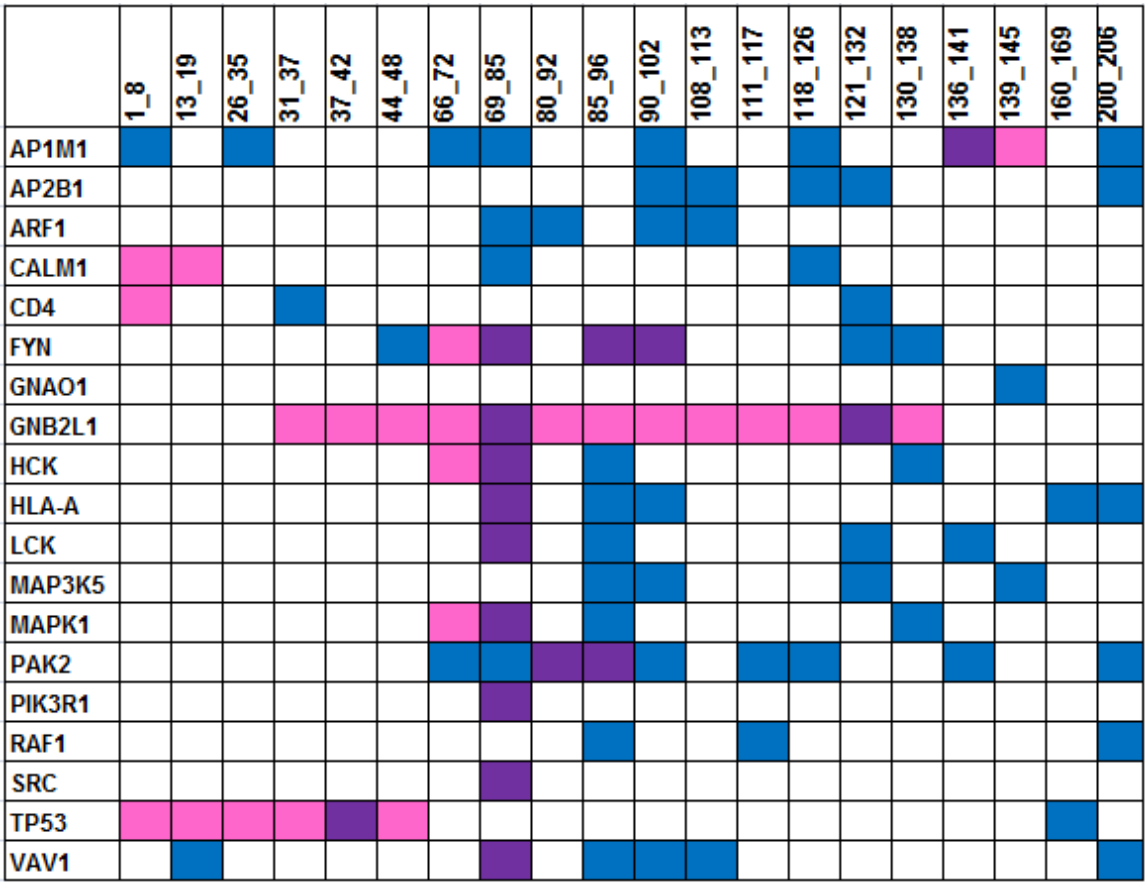

Figure 7. Hotspots and their corresponding H1 proteins

Columns indicate hotspots on Nef named by their start and end position on Nef. Rows represent gene symbols of Nef H1 proteins. An H1 protein associated with a hotspot is indicated by blue color in the corresponding box. Purple boxes are the hotspots and h1s which their interaction is validated in the literature. Pink boxes are hotspots that their interaction with the corresponding H1 protein is validated in the literature but the motif's matching the hotspot are not enriched in the neighbors of the H1 protein.

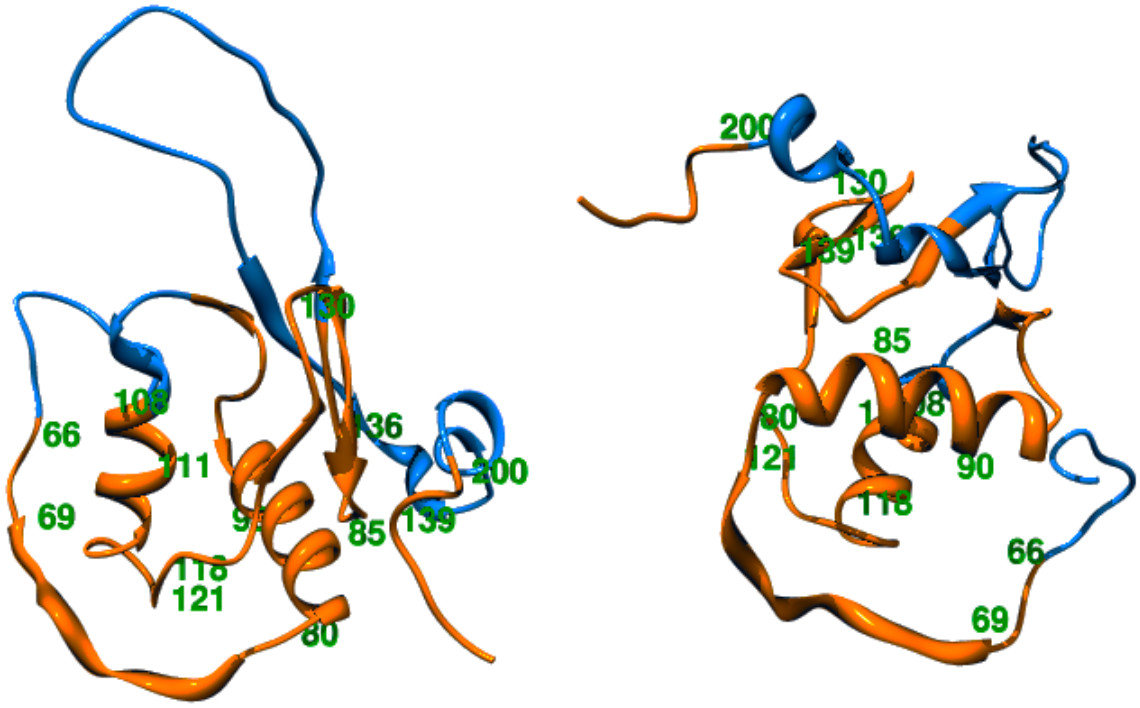

Figure 8. Motif Clusters highlighted on Nef 3D crystal structure

Regions covered by clusters of motifs highlighted in orange on Nef. PDB structure 2NEF [68] was used. Numbers on the structures reflect the start positions of the clusters. Molecular graphics images were produced using the UCSF Chimera package [69].

3.4. Discussion

In the absence of antiretroviral therapy, HIV infection causes progressive loss in CD4 lymphocyte numbers and function, resulting in the immunodeficiency associated with AIDS. Recently, HIV Nef has been shown to be the main determinant of accelerated CD4 lymphocyte depletion in vivo [106, 107]. HIV Nef is a 25- to 30-kDa myristoylated protein which is produced in the early stages of an infection [108]. In infected cells Nef localizes at the plasma membrane and preferentially associates with the cytoskeleton,

but it is also found at the nuclear membrane, in the cytoplasm, and in the nucleus [109-112]. HIV Nef has been shown to down-regulate cell surface CD4 and major histocompatibility complex class 1 molecules, augment virus infectivity, and to modulate multiple cellular signaling pathways in both CD4 lymphocytes and macrophages [113]. Nuclear localization of Nef in HIV-infected cells suggests a functional role as a nuclear regulatory factor.

Nef regulation of T-cell activation and associated pathways involves direct binding events between Nef and cellular signal transduction elements such as CD4, NAK, Raf-1, MAPK, and p53 as quantified using coprecipitation, enzyme-linked immunosorbent assay and other binding assays [114]. Similar binding assays with Nef fragments provide information about the sequence location of binding sites between Nef and host proteins [80]. For example such experiments led to the observation that HIV Nef binding to TP53 protected cells against p53-mediated apoptosis and that N terminal region of the protein up to residue 57 was responsible for binding [80].

Site-directed mutagenesis of HIV proteins and synthetic peptides from conserved regions of Nef were used to identify Nef residues crucial for binding to host HLA-A3.1 [115]. The amino acid at position 152 of the A3.1 molecule appeared to be critical for this binding event. Other site-directed mutagenesis experiments revealed the importance of Nef residues 55 to 58 in binding to CD4 [116]. Similar experiments presented evidence that a diacidic motif on Nef and the basic patch on alpha-adaptin are both required for the cooperative assembly of a CD4-Nef-AP-2 complex [80]. The research literature on

Nef binding to host proteins compiled in the HHPID database was reviewed which brought out an incomplete and perhaps a noisy portrayal of the motifs and interfaces associated with the binding of Nef to host proteins.

Lack of a high throughput binding assay for identifying Nef motifs responsible for binding to host proteins prompted me to explore the use of extensive genomics and proteomics data in a motif discovery approach to extract knowledge on Nef binding motifs targeting host proteins. Within relatively disordered regions of proteins lie short linear peptide sections that are responsible for thousands of protein–protein interactions [117]. Specific examples of motifs binding to protein domains such as the SH3 and WW domains are listed in the ELM database [27]. Peptide mediated interactions are typically transient, and occur particularly in signaling pathways [118]. The fact that Nef is relatively flexible and disordered and that it undergoes a cascade of transient interactions with the host provides strong rationale for the computational discovery of Nef peptides undergoing binding interactions with host proteins.

Although the principles of motif discovery were already established in the literature [29, 30, 32-36] and there are multiple open-access motif discovery tools that exist for short linear protein motifs, actual application to this case proved extremely challenging. First motif discovery was performed for nineteen Nef-targeted host proteins, one at a time using sets of sequences of proteins binding to the targeted host protein. This approach resulted in large numbers of motifs with no peptide instances conserved on more than a thousand Nef sequences available in the literature. Adding too many sequences of Nef

to the set of sequences for motif discovery overwhelmed the computations with peptides specific to Nef only. The optimal combination was settled to adding Nef sequences to each set at a one-to-ten ratio. Still, the output of any run of motif discovery resulted in annotation of thousands of motifs on the Nef sequence, merely over 200 residues long. Therefore, I focused only on those motifs that were relatively conserved on HIV Nef sequences, were infrequent in the host proteome but highly abundant on the binding partners of the host protein targeted by Nef.

Results showed multiple motifs projecting onto similar regions of the Nef sequence. For example, proline-rich motifs most frequently projected on Nef residues 69-96. Another hotspot was formed between residues 118 and 132. Overall, nearly half of the residues of Nef fell onto a motif discovered in this analysis. A hotspot is a sequence segment intersected by multiple motifs. The motifs were clustered based on where they fell on the Nef sequence, with the same start residue for each motif in a cluster. Clusters often intersected at the same hotspot. Next, motif similarity approaches were used to identify in some cases the ELM motifs with matching sequences satisfying both motif regular expressions. Such ELM motifs shown in Table 7 were comprised of phosphorylation site motifs for kinase substrates as well as motifs known to be used in binding to host protein domains. Comparison with known motifs presents the possibility of a functional annotation of the motif prior to experimental verification.

The results show multiple Nef hotspots for a host protein to bind to. Many of the motifs annotated in this study will likely lead to the small interfaces between Nef and its protein partners. This may result in the affinity of the interactions being weaker than that for typical interactions between globular domains with larger interfaces. Nevertheless multiple such interactions could be optimal for transient binding modes between kinases and substrates where pairs of proteins must be docked before phosphorylation occurs. How do we know which combination of Nef hotspots comprise a plausible set of binding interfaces to a given Nef targeted protein? To address this question, Nef motifs potentially targeting a given host protein were determined. Then we the binding partners of the host protein carrying the motifs under consideration were identified. The combination with the largest number of common outcompeted proteins is likely to comprise the interface of Nef with a host protein. Results shown in Table 8 require further validation. In this table, each row lists a triplet of clusters. Clusters of this triplet co-occur more than any other possible triplets for the associated Nef targeted protein (H1). Occurrences of the clusters among the neighbors of the H1 protein (Nef targeted) under consideration were calculated. For a given H1 protein and a triplet of three clusters, the last column is a probabilistic score calculated using the following formula:

$$Co - occurrence\ Score = \frac{C_{123}}{C_1 C_2 C_3}$$

Where C_{123} is the number of neighbors of H1 protein that contain all three clusters. C_1, C_2, C3 are the number of neighbors of H1 that contain the first, second and third clusters respectively. This score represents a relative likelihood of co-occurrence of the clusters in the triplet in comparison with random. For MAPK1, HCK, FYN which are all kinases, three clusters of 69 to 85, 74 to 85, and 130 to 138 occur together which suggest that not only the proline rich region of Nef (69 to 85) is important for Nef to get phosphorylated by kinases, but also another region stretching from residue 130 through 138 also potentially plays a role in the phosphorylation process.

Table 8. Co-occring clusters on Nef and outcompterd proteins

The table lists triplets of co-occuring clusters among the clusters discovered in this chapter. Cluster counts represent number of times a cluster was observed mong the neighbors of the protein associated with the cluster. Score is the relative likelihood of co-occurrence as explained in the text.

Nef Target	Cluster1	Cluster1 Count	Cluster2	Cluster2 Count	Cluster 3	Cluster 3 Count	All 3 clusters	Score
PAK2	136_141	1556	69_85	2621	200_206	1642	461	6.16
MAPK1	69_85	2465	130_138	2375	74_85	1119	403	5.51
HCK	69_85	3717	130_138	2737	74_85	2587	1113	3.78
MAP3K5	85_96	1003	139_145	2474	121_132	2464	174	2.55
FYN	69_85	4926	130_138	3870	74_85	3237	1703	2.47
AP2B1	118_126	1987	200_206	2781	108_113	2688	403	2.43
AP1M1	90_102	1267	200_206	2785	26_35	2439	224	2.33
HLA-A	163_169	1339	85_96	2736	74_85	1812	172	2.32
VAV1	69_85	4363	74_85	2676	108_113	2746	721	2.01
LCK	136_141	1339	69_85	2422	121_132	2434	177	2.01
ARF1	69_85	2860	85_96	2234	108_113	2400	340	1.98

These computations of Nef hotspots comprising binding interfaces with host proteins have implications in drug discovery. The small sizes of the interacting surfaces comprised by the motif and its counterpart makes them better candidates for intervention by small molecules than larger domain–domain interfaces [6,7]. Indeed, there are already potential drugs that inhibit protein–peptide binding. The cancer drug candidate compound Nutilin-3 disrupts the p53-MDM2 complex by mimicking a peptide in P53, one of the Nef targeted proteins. The drug is thought to free P53 to respond to DNA damage [8,9]. The research presented here could potentially identify therapeutic uses for a number of existing drugs and drugs in clinical trials

The approach taken can only uncover those Nef motifs that are also being used by host proteins to bind to interfaces on the 3D structures of host proteins, in this case, mostly kinases, transcription factors, and G proteins. If a motif is specific to a viral protein and is not part of the vocabulary of the host proteome, the approach of this study will likely not detect it correctly.

3.5. Conclusions

In this chapter, Motifs and hotspots on Nef that are associated with the binding to nineteen human proteins targeted by Nef were identified. Additionally, these hotspots were annotated with associated Nef targeted proteins. Co-operation of motifs for binding were studied by calculating co-occurrence frequencies of all of the motif clusters to identify triplets of clusters with more likelihood of co-occurence. These findings can

guide further experiments to establish the role of new binding sites which ultimately can lead to the discovery of new PPI-blocking methods in particular for HIV Nef.

Chapter 4: Connectivity Map of Iron-binding Proteins in HIV infection

4.1. Background

Metal ions play fundamental roles in switching proteins to active states in all living beings. The redox abilities of ferrous (Fe^{2+}) and ferric (Fe^{3+}) iron are essential for eukaryotic biological systems [119]. In mammals, iron in hemoglobin and myoglobin binds oxygen allowing for its transport to cells in the vicinity of blood vessels. In addition, iron acts as a cofactor for enzymes involved in energy metabolism (mitochondrial respiratory chain and Kreb's cycle) and DNA synthesis, rendering it essential for cells [120, 121]. Moreover, iron has a crucial role in immunity and immunosurveillance through involvement in cell-mediated immune effector pathways and cytokine activities. Iron promotes the growth of immune cells [120, 122, 123], thereby affecting the immune response to an invading pathogen. In return, cytokines and radicals produced and released by the immune cells control and regulate iron homeostasis via transcriptional and post-transcriptional methods [120]. Hence, iron metabolism and the immune system possess a delicate relationship through which they can regulate one another.

While iron is important for mammalian cells and deficiencies result in aberrant cell proliferation and immune function, iron overload can be deleterious [119, 120], affecting the proliferation and activation of T-cells, B-cells and natural killer cells [120, 124, 125]. One mechanism through which iron loading can affect cells is by inhibiting IFNG-

mediated pathways in macrophages, which causes them to lose their ability to kill intracellular pathogens [120]. Moreover, the lack of an iron excretory pathway in mamallian cells highlights the importance of homeostatic mechanisms adopted by cells in order to balance out iron needs as opposed to iron overload as well as redox utility as opposed to resulting toxicity.

The initial step for achieving homeostasis is through the regulation of iron absorption from the gut. However the process of transporting iron to usage and storage sites is equally important, in addition to the roles of enterocytes and macrophages [119]. Monocytes and macrophages utilize different pathways to obtain iron. These methods include transferrin-mediated uptake, transmembrane uptake of ferrous and ferric iron, obtaining iron through lactoferrin or ferritin receptors, as well as through erythrophagocytosis. As a result, the proliferation and differentiation of these cells are not affected by limiting the iron supply through one of these sources [120].

Iron-binding proteins often appear in the lists of proteins targeted by infectious agents. For example, many of the activities of the host cells targeted by HIV are iron-dependent [126]. Viruses depend on host cells for their survival and viral replication requires enhanced cellular metabolism for transcribing and translating viral genomes and proteins. Since these processes depend on and require iron, the host cells have to contain a sufficient supply of iron to meet the demands [121]. Iron accumulation can accompany the more advanced stages of HIV infection [126, 127], while increased iron

storage in bone marrow macrophages could be associated with shorter survival times [128, 129]. Elevated iron stores have been detected in other tissues of HIV patients including brain, liver and muscles [126]. In this study, public databases and bioinformatics tools were used to develop a connectivity map for HIV and host proteins in iron ion mediated signaling and metabolism. The resulting map is a portrayal of what is known about the iron ion mediated cell pathways targeted by HIV.

4.2. Methods

4.2.1. Identification of iron-associated proteins

To determine the human proteins that are associated with iron binding, a list of relevant Gene Ontology [130] molecular function categories were obtained. List of proteins annotated with GO molecular function *iron ion binding* were retrieved from the GO Consortium [130] and DAVID Bioinformatics [104]. For each category the two protein list were compared and those found only in DAVID were checked against the literature and UniProtKB [131] database to confirm their functional association with iron. Proteins with literature support were merged to produce a final list for each category. Only genes with RefSeq status of "REVIEWED" or "VALIDATED" were retained; in this process 18 genes were eliminated.

In addition to the proteins annotated with the aforementioned GO categories, 6 others were added to the list of iron-associated proteins based on a review paper by Drakesmith and Prientice on iron metabolism and viral infection [121]. GLRX2 was

among the iron binding proteins and GLRX5 which is a protein from the same family is annotated with iron-sulfur cluster binding were added to the list as well. ERCC2 is an iron-sulfur complex binding protein which is part of a complex that binds TAT and is part of the HIV-1 transcription regulatory process therefore it was added to the list. The final list contained 299 proteins associated with iron.

4.2.2. Pathway visualization

A list of proteins interacting with HIV-1 proteins was obtained from the NIAID HIV-1, Human Protein Interaction Database [52] (Version: December 2009) containing 1433 different human proteins with 68 types of interactions (e.g. binds, activates, upregulates, downregulates, etc.). The overlap between this list and the list of iron associated proteins resulted in 40 proteins associated with 99 different interactions with 12 HIV-1 proteins. CYP27B1 was the only protein whose interaction with the HIV-1 matrix protein was ambiguous and did not fit any of the pathway notations so it was removed from the pathway. CellDesigner [132] was used to visualize these interactions between the HIV-1 proteins and the human iron-associated proteins in a pathway based on interaction types and intracellular locations of the proteins involved. For each interaction an extensive PubMed search was conducted to determine type, intracellular location and conditions under which the interactions take place. These literature searches allowed for the addition of other essential human proteins associated with the main interactions in the pathway. PubMed IDs of relevant papers of each interaction were added as notes to provide quick access through CellDesigner. SMBL notation available in CellDesigner

was adapted to visualize these interactions and make changes when needed. The list of notations and symbols used in the pathway is shown in Figure 10.

4.3. Results and Discussion

A total of 313 proteins were identified to be iron-associated with the use of gene ontology categories. Out of 313 proteins, the iron binding motif [DE]..E is present on 289 of them. However, the motif is ubiquitously present on the human proteome; 30466 out of 38829 NCBI human proteins express the motif. Therefore, the presence of the motif in its current form was not used to further restrict the subset of host iron-binding proteins.

Among the host iron-binding proteins, 40 appeared in the HIV-1, Human Protein Interaction Database (HHPID) as a subset of the 1393 host proteins targeted by HIV (Table 9). A hypergreometric test was used to calculate the significance of the overlap using all human proteins from NCBI as the background. The p-value was found to be 1.8e-12, indicating a high-level of statistical enrichment of iron-associated proteins among known HIV-1 interacting host proteins. Iron binding host proteins targeted by HIV populate different modes of crosstalk with virus proteins as shown in Table 9. Iron-binding proteins appear more often in HHPID as 'upregulated' and 'stimulated by' than in direct 'binding' interactions. The reason these proteins may appear more affected by HIV infection than the rest of the human protein pool is because they are more centrally connected in the protein network as assessed by estimation of their binding partners in the human protein network. Nevertheless, a system view, as adopted here, may help

construct links among the research results concerning different regions of the connectivity map developed in this study.

The connectivity map shown in Figure 10 presents a network scale of interactions of HIV proteins with iron binding proteins of the host. The map uses a notation described in Figure 11 that identifies the modes of interaction. The map also includes host proteins neither interacting with iron ions nor with HIV but have central connectivity roles in the map. In the paragraphs below the important loci of the connectivity map are highlighted. These proteins are not necessarily among the ones directly interacting with HIV proteins but are centrally connected in the map.

Table 9. Human Iron binding proteins- HIV-1 interactions

Types of interactions between HIV-1 proteins and iron-associated proteins

SYMBOL	Gene ID	Interaction Type	HIV
ABCE1	6059	associates with	PR55
		associates with	VIF
ALOX5	240	upregulated by	GP120
APP	351	activated by	RETROPEPSIN
		inhibited by	GP41
		inhibits	GP120
		inhibits	TAT
		upregulated by	TAT
CAT	847	inhibits	GP160
CYBB	1536	inhibited by	CASPID
CYC1	1537	release induced by	VPR
CYCS	54205	released by	VPR
CYP51A1	1595	upregulated by	NEF
DOCK2	1794	associates with	NEF
GLRX2	51022	activates	RETROPEPSIN
GLRX5	51218	activates	RETROPEPSIN
HFE	3077	downregulated by	NEF
HMOX2	3163	upregulated by	GP120
IDO1	3620	release induced by	GP120

Table 9. (continued)

SYMBOL	Gene ID	Interaction Type	HIV
IKK1	1147	binds	GP120
		phosphorylated by	NEF
IKK2	3551	binds	GP120
		phosphorylated by	NEF
IKKE	9641	binds	GP120
		phosphorylated by	NEF
LTF	4057	inhibits	GP120
NOS1	4842	inhibited by	TAT
		upregulated by	GP41
NOS2	4843	inhibited by	TAT
		upregulated by	GP120
		upregulated by	gp41
NOS3	4846	inhibited by	TAT
		upregulated by	gp41
NOX1	27035	activated by	GP120
NOX3	50508	activated by	GP120
NOX4	50507	activated by	GP120
NOX5	79400	activated by	GP120
PPP1CA	5499	downregulated by	GP120
		stimulates	TAT
PPP1CB	5500	stimulates	TAT
		upregulated by	GP120
PPP1CC	5501	stimulates	TAT
PPP2CA	5515	inhibits	TAT
PPP2CB	5516	inhibits	TAT
PPP3CA	5530	activated by	TAT
PPP3CB	5532	activated by	TAT
PPP3CC	5533	activated by	TAT
PTGS1	5742	upregulated by	GP120
		upregulated by	TAT
PTGS2	5743	upregulated by	GP120
		upregulated by	TAT
SDHB	6390	binds	TAT
TFRC	7037	downregulated by	GP120
		downregulated by	NEF
TH	7054	downregulated by	TAT

NADPH-Oxidase: NADPH oxidase is an enzymatic complex composed of multiple proteins. Iron is essential for the functioning of the NADPH oxidase complex with a heme-b acting as the prosthetic redox group in cytochrome b. Iron deficiencies therefore

result in reduced enzyme activity [133]. NADPH oxidase is the main producer of superoxide anion (O_2^-) through the reduction of oxygen. In the cell, superoxide dismutase (SOD) then acts as an antioxidant by utilizing electrons from copper or zinc for the conversion of superoxide into hydrogen peroxide (H_2O_2). In resting cells, the NADPH oxidase complex is typically dormant. Monocytes and macrophages usually release increased levels of reactive oxygen species (ROS) as a response to certain stimuli. The generation of high levels of ROS, referred to as a respiratory burst, plays an important role in the host defense mechanism against pathogens [134, 135]. These reactive species are therefore involved in inflammatory processes, apoptosis, aging and carcinogenesis [136].

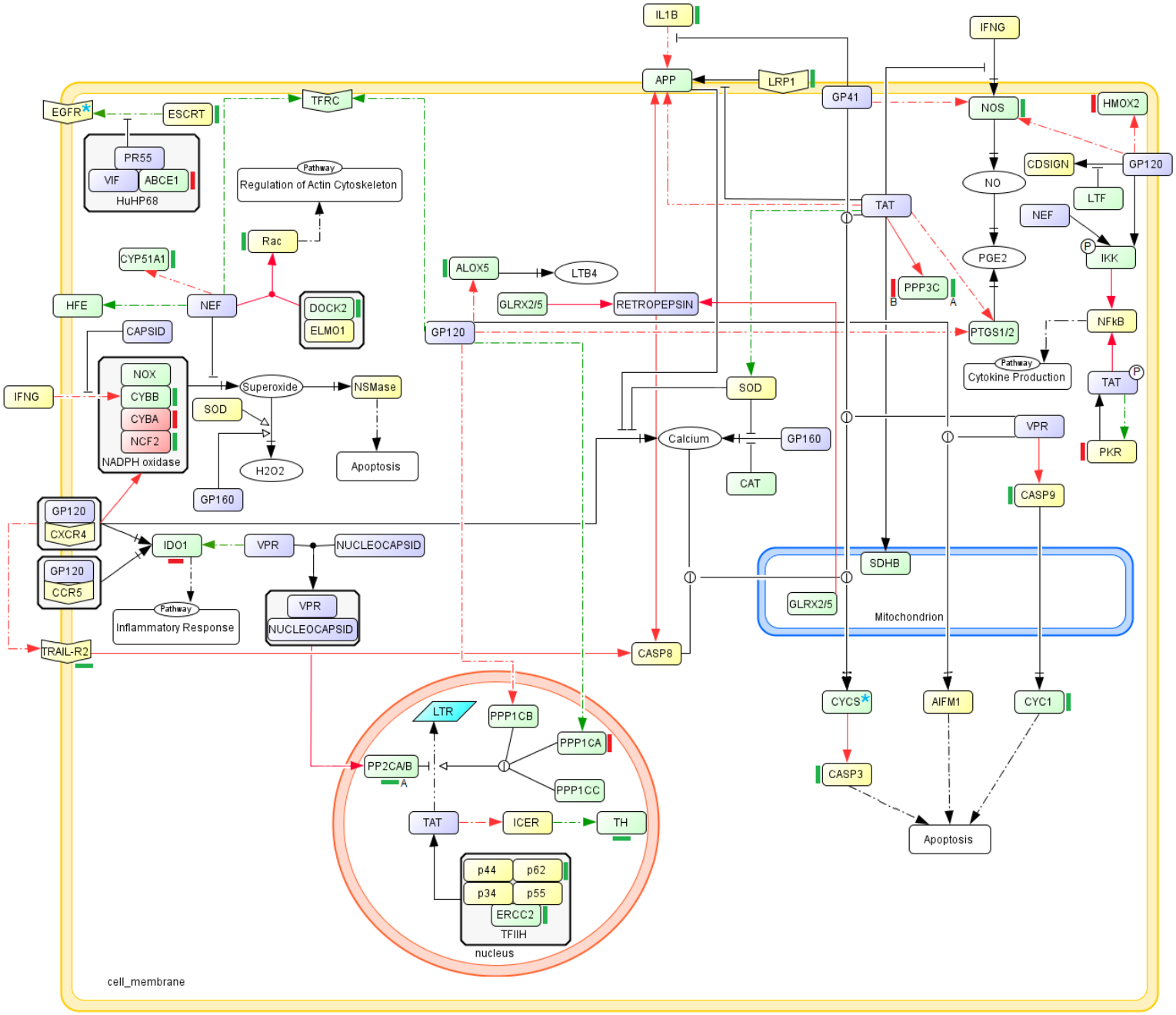

Figure 9. Pathway of interactions between iron binding proteins and HIV-1 proteins

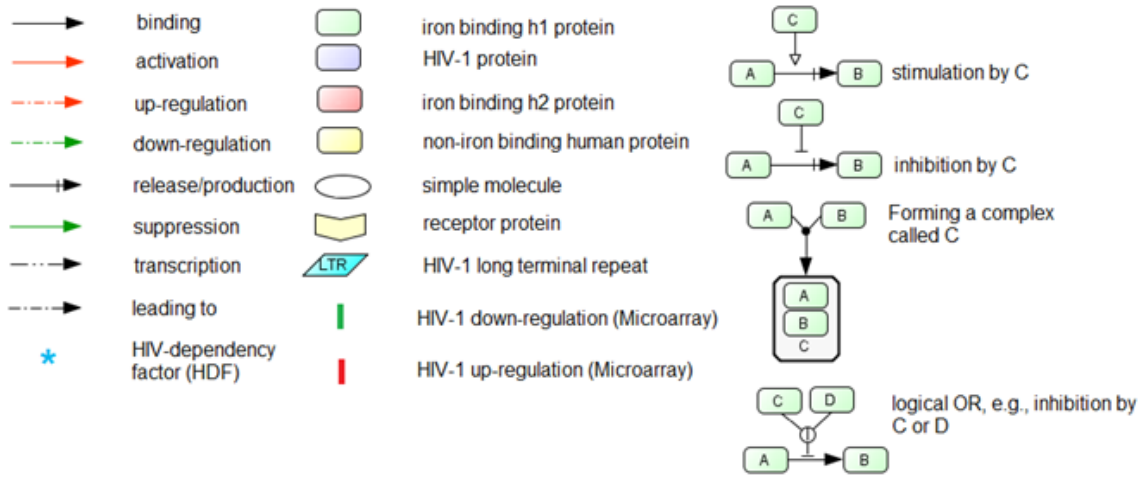

Figure 10. Pathway Notations
Legend showing different notations used in the pathway

HIV-1 targets NADPH oxidase indirectly, through other proteins. First, gp120 binds to CXC chemokine receptor 4 (CXCR4) which in turn activates the NAPDH oxidase complex resulting in increased expression of superoxide radicals and subsequent activation of neutral sphingomyelinase, inducing apoptosis and cell death [137]. On the other hand Nef plays a time-dependent role in this process. In the early stages, Nef is responsible for the induction of phosphorylation and cell-membrane translocation of NCF1 and NCF2, hence activating NADPH oxidase, which results in the production of superoxide [134, 135]. Meanwhile, gp160 also enhances the respiratory burst and oxidative stress through the production of H_2O_2 [136]. Within 10 hours however, Nef inhibits NADPH oxidase resulting in a dysregulation in the production of ROS, impairing specific immune functions including the oxidative burst response and

phagocytosis. This in turn allows for the development of HIV-1 pathogenesis [134, 135]. In addition, the viral capsid has been shown to inhibit the interferon-gamma (IFN-γ) induced accumulation of the cytochrome B heavy chain mRNA, which is a component of the NADPH oxidase complex [138].

HuHP68 (ABCE1) Complex: While the HIV protein Vif is excluded from the mature viral particles, it is essential for viral infectivity. It is therefore a late HIV-1 product, acting in the latter stages of the virus life cycle during viral assembly and/or maturation to enhance the infectivity of the progeny virions [139, 140]. Vif interacts with cellular ABCE1 and viral PR55 (Gag) to assist in capsid assembly. ABCE1 is known to function as an RNAse L inhibitor, suggesting that the viral association with ABCE1 is possibly to protect the viral RNA from degradation during viral assembly [139]. HIV-1 Gag polypeptides are synthesized in the cytoplasm of infected cells and then are trafficked to the plasma membrane. ABCE1 is then recruited to sites of assembling Gag at the membrane and the association continues throughout capsid formation until the onset of viral maturation and its subsequent release [141]. Typically ABCE1 is required for cellular survival, mRNA translation, and ribosome biogenesis. It is the only ATP-binding cassette enzyme that has an amino-terminal iron-sulfur cluster domain, thus necessitating the availability of iron for its functioning [121, 142].

In addition to ABCE1, PR55 also recruits the ESCRT (endosomal sorting complex required for transport) pathway protein VPS23 (TSG101) from the ESCRT-I complex as

well as PDCD6IP which interacts with ESCRT-I and CHMP4 (from ESCRT-III) proteins [143, 144]. ESCRT proteins are usually involved in the sorting of ubiquinated proteins including ligand-activated cell surface receptors in order to deliver them into the lumens of multivesicular bodies. VPS23 typically results in the degradation of the epidermal growth factor receptor (EGFR). However as Gag binds to and hence depletes VPS23, the rate of EGFR down-regulation is reduced resulting in increased intracellular retention of EGFR. This in turn allows for prolonged EGFR-mediated signaling through ERK/MAP kinase and hence cellular proliferation [144].

Indoleamine 2,3-dioxygenase 1 (IDO1): IDO1 in the connectivity map contains a heme-prosthetic group in its center. The iron is present in the ferric (Fe^{3+}) form in the inactive state and as Fe^{2+} in the active state [145]. IDO1 catalyzes the degradation of the essential amino acid L-tryptophan to N-formyl-kynurenine by incorporating molecular oxygen or a superoxide anion [146]. Superoxides are produced in high quantities at sites of infection or inflammation and IDO1 is known to be involved during the innate immune response of the host [145]. It is therefore an immunosuppressive enzyme that results in the suppression of T cell proliferation through the catabolism of tryptophan [147]. The increased expression of IDO1 and its down-stream metabolites of kynurenine is associated with several central nervous system disorders, including AIDS dementia complex [146]. During HIV-1 infection, IDO1 is regulated through both GP120 and Vpr viral proteins, which possess antagonistic effects on the cellular expression of IDO1. GP120 interacts with the cell surface coreceptors chemokine (C-C motif) receptor 4

(CCR4) and CXCR5. It then induces increased production of IDO mRNA levels resulting in loss of CD4+ T cell function [147]. Vpr, on the other hand, results in increased glucocorticoids and in turn it affects the gene expression of glucocorticoid-regulated genes including the down-regulation of IDO1 [148].

CYP51A1 Cholesterol is necessary for HIV-1's entry into and budding out of the host cell. Since Nef is highly expressed in the early stages of the replication cycle of HIV-1, it can help facilitate the viral entry into the cell. Nef is believed to influence cholesterol production by increasing the expression of the cytochrome P450 51 (CYP51) gene in cDNA microarray tests in Jurkat cells [149] CYP51 is responsible for the 14α demethylation of lanosterol, necessary for ersgosterol biosynthesis [150]. Like other cytochrome P450 family members, CYP51 contains a heme iron in its active site [151]. Blockage of the ersgosterol biosynthesis results in impairments of the membrane integrity and the function of membrane-associated proteins [150].

ALOX5: Arachidonate 5-lipoxygenase (ALOX5) is a nonheme iron-containing dioxygenase that plays an important role in the biosynthesis of leukotrienes, namely the catalysis of the production of leukotriene LTA4 from arachidonic acid, which can then be converted to LTB4 [152]. Leukotrienes are important inflammatory mediators and LTB4 can then induce the adhesion and activation of leukocytes, ALOX5 is therefore mainly expressed in the different leukocytes [153]. Research by Maccarrone et al. [154, 155] suggests that the HIV-1 coat glycoprotein GP120 enhances the activity of ALOX5

resulting in increased LTB4 production and consequent cell death in neuroblastoma cells. In addition, ALOX5 might be capable of inducing cell cytotoxicity by oxidizing cellular membranes [154, 155]. Nonetheless, leukotriene synthesis is in fact reduced in the macrophages and peripheral mononuclear cells of HIV patients [156-158], although GP120 was not cited as the cause.

Figure 11 is a bar graph comparing interaction types between all interactions in the HIV-1 Human Interaction Database and those interactions that involve an iron binding protein. The graph suggests that HIV utilizes iron binding proteins more in regulatory alterations including up/down regulations or activation.

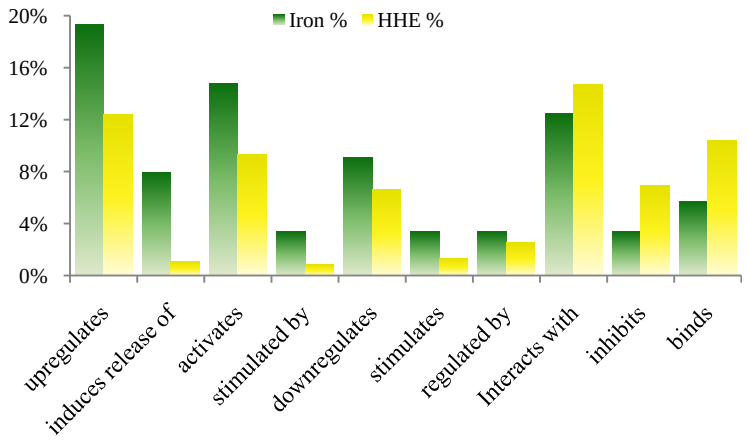

Figure 11. Frequency of different types of HIV-1 interaction types

The chart reflects percentage of top 10 interaction types between iron-binding proteins and HIV-1 proteins (green bars) in compare with all interactions of HIV-1 proteins with human proteins (yellow bars)

Overall the map indicates loci of human protein networks important for HIV infection. The map would be strengthened by the identification of regions within it that are important for other viral infections. The set of proteins appearing in the map can be used to test how imposed changes on cells can alter iron ion binding protein network in gene set enrichment analysis through microarray experiments.

4.4. Conclusions

In this study, a connectivity map uncovering crosstalk between HIV and the iron ion mediated signaling and metabolism pathways of the host was developed. This notation is detailed enough to summarize the research findings with accuracy. The connectivity map thus produced, when integrated with further network analysis, will guide researchers to the impact of viral infections on iron ion dependent processes in the host cell.

Chapter 5: Conclusions

This doctoral thesis focused on the use of bioinformatics databases and tools in the discovery of the biological rules governing HIV-host protein interactions. For this purpose, we used protein-sequence data, large-scale data on human protein binding interactions, a database on HIV-1, Human Protein Interactions, along with extensive research literature. The bioinformatics tools utilized included statistical enrichment analysis, motif discovery methods, gene ontology and pathway analysis, network development tools, and code writing. This work is novel in the system level approach to motif discovery in combinations of host and viral proteins and interpreting the complex data resulting from this approach. The viral motif-host protein interactions presented in my thesis comprise important contributions to research literature on HIV-host crosstalk. In addition, through the use of iron-ion dependent host cell mechanisms, it has been illustrated how currently available network building techniques enable one to integrate patchy research literature into a portrayal of species crosstalk affecting modes of host protein networks.

Motifs and hotspots discovered and annotated in Chapters 2 and 3 can be further studied by projecting them to available 3D structures of proteins and calculating their surface availability. Also co-operative binding of such hotspots can be verified by experimental methods. Emerging soft docking algorithms along with increasing

numbers of available protein structures in PDB will allow for the discovery of hotspots and motifs counterpart.

Main conclusions reached in this thesis are as follows:

- HIV virus proteins express sequential hotspots for interacting with host hub proteins. These hotspots intersect with multiple viral motifs indicating how diversity in virus-host proteins is achieved through the use of a finite set of sequence hotspots in crosstalk with the host.
- HIV sequence hotspots predicted in the present study accurately produce much of the research literature on sites of binding events of HIV proteins.
- Predictions presented in this thesis could be further studied by experimental approaches involving viral protein segments and site-directed mutagenesis experiments.
- The study on HIV Nef presented in Chapter 3, uncovers motifs and hotspots potentially used by this virus protein to interact with nineteen host proteins.
- The study presented in this dissertation, indicates the presence of multiple viral hotspots for binding to a given host protein, suggesting that motif-host protein combinations need to be considered as a dominant mode of interaction in virus-host protein binding interactions.
- This thesis exposes the many challenges of current motif discovery approaches in their application to the discovery of the grammar of crosstalk between species.

Current methods produce way too many motifs that could potentially be integrated without loss of specificity. The methods become more limited in their efficiency with increasing numbers of protein sequences for motif discovery.

- Results presented in this dissertation, produce candidate sites for further experimental sites and provide rationale for developing high throughput experimental schemes for virus host binding experiments.
- The study presented in this thesis is relevant to current treatments of HIV infection as it suggests viral sequence sites to be targeted by current and newly developing drug regiments. Moreover, the methodology developed in this thesis, can also be used to study other virus-host interactions.
- In this work it was shown how network building techniques can be used to draw portrays of host-virus crosstalk. Illustrated in this case for HIV – iron ion binding proteins.
- This thesis underscores the need for further development of bioinformatics tools for binding interface discovery including 3D docking simulations and molecular dynamics simulations.

List of References

1. Ansorge WJ: **Next-generation DNA sequencing techniques**. *N Biotechnol* 2009, **25**(4):195-203.
2. Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y: **A comprehensive two-hybrid analysis to explore the yeast protein interactome**. *Proc Natl Acad Sci U S A* 2001, **98**(8):4569-4574.
3. Uetz P, Hughes RE: **Systematic and large-scale two-hybrid screens**. *Curr Opin Microbiol* 2000, **3**(3):303-308.
4. Zhu H, Bilgin M, Bangham R, Hall D, Casamayor A, Bertone P, Lan N, Jansen R, Bidlingmaier S, Houfek T *et al*: **Global analysis of protein activities using proteome chips**. *Science* 2001, **293**(5537):2101-2105.
5. Rual JF, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, Li N, Berriz GF, Gibbons FD, Dreze M, Ayivi-Guedehoussou N *et al*: **Towards a proteome-scale map of the human protein-protein interaction network**. *Nature* 2005, **437**(7062):1173-1178.
6. Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck FH, Goehler H, Stroedicke M, Zenkner M, Schoenherr A, Koeppen S *et al*: **A human protein-protein interaction network: a resource for annotating the proteome**. *Cell* 2005, **122**(6):957-968.
7. Dziembowski A, Seraphin B: **Recent developments in the analysis of protein complexes**. *FEBS Lett* 2004, **556**(1-3):1-6.
8. Diella F, Haslam N, Chica C, Budd A, Michael S, Brown NP, Trave G, Gibson TJ: **Understanding eukaryotic linear motifs and their role in cell signaling and regulation**. *Front Biosci* 2008, **13**:6580-6603.
9. Dunker AK, Silman I, Uversky VN, Sussman JL: **Function and structure of inherently disordered proteins**. *Curr Opin Struct Biol* 2008, **18**(6):756-764.
10. Stein A, Pache RA, Bernado P, Pons M, Aloy P: **Dynamic interactions of proteins in complex networks: a more structured view**. *FEBS J* 2009, **276**(19):5390-5405.
11. Arhel N, Kirchhoff F: **Host proteins involved in HIV infection: new therapeutic targets**. *Biochim Biophys Acta* 2010, **1802**(3):313-321.
12. Betzi S, Restouin A, Opi S, Arold ST, Parrot I, Guerlesquin F, Morelli X, Collette Y: **Protein protein interaction inhibition (2P2I) combining high throughput and virtual screening: Application to the HIV-1 Nef protein**. *Proc Natl Acad Sci U S A* 2007, **104**(49):19256-19261.
13. Haffar O, Dubrovsky L, Lowe R, Berro R, Kashanchi F, Godden J, Vanpouille C, Bajorath J, Bukrinsky M: **Oxadiazols: a new class of rationally designed anti-**

human immunodeficiency virus compounds targeting the nuclear localization signal of the viral matrix protein. *J Virol* 2005, **79**(20):13028-13036.

14. He Y, Cheng J, Li J, Qi Z, Lu H, Dong M, Jiang S, Dai Q: **Identification of a critical motif for the human immunodeficiency virus type 1 (HIV-1) gp41 core structure: implications for designing novel anti-HIV fusion inhibitors**. *J Virol* 2008, **82**(13):6349-6358.
15. Goh CS, Milburn D, Gerstein M: **Conformational changes associated with protein-protein interactions**. *Curr Opin Struct Biol* 2004, **14**(1):104-109.
16. Stein A, Aloy P: **Contextual specificity in peptide-mediated protein interactions**. *PLoS One* 2008, **3**(7):e2524.
17. Berman HM, Battistuz T, Bhat TN, Bluhm WF, Bourne PE, Burkhardt K, Feng Z, Gilliland GL, Iype L, Jain S *et al*: **The Protein Data Bank**. *Acta Crystallogr D Biol Crystallogr* 2002, **58**(Pt 6 No 1):899-907.
18. McEntyre JR, Gibson TJ: **Patterns and clusters within the PSM column in TiBS, 1992-2004**. *Trends Biochem Sci* 2004, **29**(12):627-633.
19. Pawson T, Linding R: **Synthetic modular systems--reverse engineering of signal transduction**. *FEBS Lett* 2005, **579**(8):1808-1814.
20. Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, Telikicherla D, Raju R, Shafreen B, Venugopal A *et al*: **Human Protein Reference Database--2009 update**. *Nucleic Acids Res* 2009, **37**(Database issue):D767-772.
21. Mathivanan S, Periaswamy B, Gandhi TK, Kandasamy K, Suresh S, Mohmood R, Ramachandra YL, Pandey A: **An evaluation of human protein-protein interaction data in the public domain**. *BMC Bioinformatics* 2006, **7 Suppl 5**:S19.
22. Fuxreiter M, Tompa P, Simon I: **Local structural disorder imparts plasticity on linear motifs**. *Bioinformatics* 2007, **23**(8):950-956.
23. Hunt T: **Protein sequence motifs involved in recognition and targeting: a new series**. *Trends in Biological Sciences* 1990, **15**:305.
24. Dinkel H, Sticht H: **A computational strategy for the prediction of functional linear peptide motifs in proteins**. *Bioinformatics* 2007, **23**(24):3297-3303.
25. Mayer BJ: **SH3 domains: complexity in moderation**. *J Cell Sci* 2001, **114**(Pt 7):1253-1263.
26. Li SS: **Specificity and versatility of SH3 and other proline-recognition domains: structural basis and implications for cellular signal transduction**. *Biochem J* 2005, **390**(Pt 3):641-653.
27. Gould CM, Diella F, Via A, Puntervoll P, Gemund C, Chabanis-Davidson S, Michael S, Sayadi A, Bryne JC, Chica C *et al*: **ELM: the status of the 2010 eukaryotic linear motif resource**. *Nucleic Acids Res* 2010, **38**(Database issue):D167-180.

28. Neduva V, Russell RB: **Peptides mediating interaction networks: new leads at last**. *Curr Opin Biotechnol* 2006, **17**(5):465-471.
29. Davey NE, Edwards RJ, Shields DC: **The SLiMDisc server: short, linear motif discovery in proteins**. *Nucleic Acids Res* 2007, **35**(Web Server issue):W455-459.
30. Edwards RJ, Davey NE, Shields DC: **SLiMFinder: a probabilistic method for identifying over-represented, convergently evolved, short linear motifs in proteins**. *PLoS ONE* 2007, **2**(10):e967.
31. Henschel A, Kim WK, Schroeder M: **Equivalent binding sites reveal convergently evolved interaction motifs**. *Bioinformatics* 2006, **22**(5):550-555.
32. Li H, Li J, Wong L: **Discovering motif pairs at interaction sites from protein sequences on a proteome-wide scale**. *Bioinformatics* 2006, **22**(8):989-996.
33. Neduva V, Russell RB: **DILIMOT: discovery of linear motifs in proteins**. *Nucleic Acids Res* 2006, **34**(Web Server issue):W350-355.
34. Tan SH, Hugo W, Sung WK, Ng SK: **A correlated motif approach for finding short linear motifs from protein interaction networks**. *BMC Bioinformatics* 2006, **7**:502.
35. Neduva V, Linding R, Su-Angrand I, Stark A, de Masi F, Gibson TJ, Lewis J, Serrano L, Russell RB: **Systematic discovery of new recognition peptides mediating protein interaction networks**. *PLoS Biol* 2005, **3**(12):e405.
36. Rigoutsos I, Floratos A: **Combinatorial pattern discovery in biological sequences: The TEIRESIAS algorithm**. *Bioinformatics* 1998, **14**(1):55-67.
37. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs**. *Nucleic Acids Res* 1997, **25**(17):3389-3402.
38. Dickerson JE, Pinney JW, Robertson DL: **The biological context of HIV-1 host interactions reveals subtle insights into a system hijack**. *BMC Syst Biol* 2010, **4**:80.
39. Kadaveru K, Vyas J, Schiller MR: **Viral infection and human disease--insights from minimotifs**. *Front Biosci* 2008, **13**:6455-6471.
40. Evans P, Dampier W, Ungar L, Tozeren A: **Prediction of HIV-1 virus-host protein interactions using virus and host sequence motifs**. *BMC Med Genomics* 2009, **2**:27.
41. Davis FP, Barkan DT, Eswar N, McKerrow JH, Sali A: **Host pathogen protein interactions predicted by comparative modeling**. *Protein Sci* 2007, **16**(12):2585-2596.
42. Dampier W, Evans P, Ungar L, Tozeren A: **Host sequence motifs shared by HIV predict response to antiretroviral therapy**. *BMC Med Genomics* 2009, **2**:47.
43. Mujawar Z, Rose H, Morrow MP, Pushkarsky T, Dubrovsky L, Mukhamedova N, Fu Y, Dart A, Orenstein JM, Bobryshev YV *et al*: **Human immunodeficiency**

virus impairs reverse cholesterol transport from macrophages. *PLoS Biol* 2006, **4**(11):e365.

44. Mueller SM, Lang SM: **The first HxRxG motif in simian immunodeficiency virus mac239 Vpr is crucial for G(2)/M cell cycle arrest**. *J Virol* 2002, **76**(22):11704-11709.
45. Martins MA, Wilson NA, Reed JS, Ahn CD, Klimentidis YC, Allison DB, Watkins DI: **T-cell correlates of vaccine efficacy after a heterologous simian immunodeficiency virus challenge**. *J Virol* 2010, **84**(9):4352-4365.
46. Solorzano A, Ye J, Perez DR: **Alternative live-attenuated influenza vaccines based on modifications in the polymerase genes protect against epidemic and pandemic flu**. *J Virol* 2010, **84**(9):4587-4596.
47. Dyer MD, Murali TM, Sobral BW: **The landscape of human proteins interacting with viruses and other pathogens**. *PLoS Pathog* 2008, **4**(2):e32.
48. Tastan O, Qi Y, Carbonell JG, Klein-Seetharaman J: **Prediction of interactions between HIV-1 and human proteins by information integration**. *Pac Symp Biocomput* 2009:516-527.
49. Ekman D, Light S, Bjorklund AK, Elofsson A: **What properties characterize the hub proteins of the protein-protein interaction network of Saccharomyces cerevisiae?** *Genome Biol* 2006, **7**(6):R45.
50. Liu Y, Tozeren A: **Modular composition predicts kinase/substrate interactions**. *BMC Bioinformatics* 2010, **11**(1):349.
51. Evans P, Sacan A, Ungar L, Tozeren A: **Sequence alignment reveals possible MAPK docking motifs on HIV proteins**. *PLoS One* 2010, **5**(1):e8942.
52. Fu W, Sanders-Beer BE, Katz KS, Maglott DR, Pruitt KD, Ptak RG: **Human immunodeficiency virus type 1, human protein interaction database at NCBI**. *Nucleic Acids Res* 2009, **37**(Database issue):D417-422.
53. Balakrishnan S, Tastan O, Carbonell J, Klein-Seetharaman J: **Alternative paths in HIV-1 targeted human signal transduction pathways**. *BMC Genomics* 2009, **10 Suppl 3**:S30.
54. Harada K, Ishida Y: **A hub gene in an HIV-1 gene regulatory network is a promising target for anti-HIV-1 drugs**. *Artificial Life and Robotics* 2009, **14**:4.
55. Neduva V, Russell RB: **Linear motifs: evolutionary interaction switches**. *FEBS Lett* 2005, **579**(15):3342-3345.
56. Ackerson B, Rey O, Canon J, Krogstad P: **Cells with high cyclophilin A content support replication of human immunodeficiency virus type 1 Gag mutants with decreased ability to incorporate cyclophilin A**. *J Virol* 1998, **72**(1):303-308.
57. Hiipakka M, Poikonen K, Saksela K: **SH3 domains with high affinity and engineered ligand specificities targeted to HIV-1 Nef**. *J Mol Biol* 1999, **293**(5):1097-1106.

58. Craig HM, Pandori MW, Riggs NL, Richman DD, Guatelli JC: **Analysis of the SH3-binding region of HIV-1 nef: partial functional defects introduced by mutations in the polyproline helix and the hydrophobic pocket**. *Virology* 1999, **262**(1):55-63.

59. Wang H, Zhang HM, Jiang Q, Peng QL, Tan Y, Li TS, Zhou BP: **[Evolution of HIV-1 drug resistance in patients failing combination antiretroviral therapy]**. *Zhonghua Yi Xue Za Zhi* 2010, **90**(9):584-587.

60. Beauparlant P, Kwon H, Clarke M, Lin R, Sonenberg N, Wainberg M, Hiscott J: **Transdominant mutants of I kappa B alpha block Tat-tumor necrosis factor synergistic activation of human immunodeficiency virus type 1 gene expression and virus multiplication**. *J Virol* 1996, **70**(9):5777-5785.

61. Ammosova T, Berro R, Jerebtsova M, Jackson A, Charles S, Klase Z, Southerland W, Gordeuk VR, Kashanchi F, Nekhai S: **Phosphorylation of HIV-1 Tat by CDK2 in HIV-1 transcription**. *Retrovirology* 2006, **3**:78.

62. Yang X, Goncalves J, Gabuzda D: **Phosphorylation of Vif and its role in HIV-1 replication**. *J Biol Chem* 1996, **271**(17):10121-10129.

63. Jian H, Zhao LJ: **Pro-apoptotic activity of HIV-1 auxiliary regulatory protein Vpr is subtype-dependent and potently enhanced by nonconservative changes of the leucine residue at position 64**. *J Biol Chem* 2003, **278**(45):44326-44330.

64. Nie Z, Bergeron D, Subbramanian RA, Yao XJ, Checroune F, Rougeau N, Cohen EA: **The putative alpha helix 2 of human immunodeficiency virus type 1 Vpr contains a determinant which is responsible for the nuclear translocation of proviral DNA in growth-arrested cells**. *J Virol* 1998, **72**(5):4104-4115.

65. Schindler M, Rajan D, Banning C, Wimmer P, Koppensteiner H, Iwanski A, Specht A, Sauter D, Dobner T, Kirchhoff F: **Vpu serine 52 dependent counteraction of tetherin is required for HIV-1 replication in macrophages, but not in ex vivo human lymphoid tissue**. *Retrovirology* 2010, **7**:1.

66. Bayer P, Kraft M, Ejchart A, Westendorp M, Frank R, Rosch P: **Structural studies of HIV-1 Tat protein**. *J Mol Biol* 1995, **247**(4):529-535.

67. Dimattia MA, Watts NR, Stahl SJ, Rader C, Wingfield PT, Stuart DI, Steven AC, Grimes JM: **Implications of the HIV-1 Rev dimer structure at 3.2 A resolution for multimeric binding to the Rev response element**. *Proc Natl Acad Sci U S A* 2010, **107**(13):5810-5814.

68. Grzesiek S, Bax A, Hu JS, Kaufman J, Palmer I, Stahl SJ, Tjandra N, Wingfield PT: **Refined solution structure and backbone dynamics of HIV-1 Nef**. *Protein Sci* 1997, **6**(6):1248-1263.

69. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE: **UCSF Chimera--a visualization system for exploratory research and analysis**. *J Comput Chem* 2004, **25**(13):1605-1612.

70. Srinivas SK, Srinivas RV, Anantharamaiah GM, Compans RW, Segrest JP: **Cytosolic domain of the human immunodeficiency virus envelope glycoproteins binds to calmodulin and inhibits calmodulin-regulated proteins**. *J Biol Chem* 1993, **268**(30):22895-22899.

71. Krogstad P, Geng YZ, Rey O, Canon J, Ibarrondo FJ, Ackerson B, Patel J, Aldovini A: **Human immunodeficiency virus nucleocapsid protein polymorphisms modulate the infectivity of RNA packaging mutants**. *Virology* 2002, **294**(2):282-288.

72. Burnette B, Yu G, Felsted RL: **Phosphorylation of HIV-1 gag proteins by protein kinase C**. *J Biol Chem* 1993, **268**(12):8698-8703.

73. Matsubara M, Jing T, Kawamura K, Shimojo N, Titani K, Hashimoto K, Hayashi N: **Myristoyl moiety of HIV Nef is involved in regulation of the interaction with calmodulin in vivo**. *Protein Sci* 2005, **14**(2):494-503.

74. Saksela K, Cheng G, Baltimore D: **Proline-rich (PxxP) motifs in HIV-1 Nef bind to SH3 domains of a subset of Src kinases and are required for the enhanced growth of Nef+ viruses but not for down-regulation of CD4**. *EMBO J* 1995, **14**(3):484-491.

75. Greenway A, Azad A, Mills J, McPhee D: **Human immunodeficiency virus type 1 Nef binds directly to Lck and mitogen-activated protein kinase, inhibiting kinase activity**. *J Virol* 1996, **70**(10):6701-6708.

76. Linnemann T, Zheng YH, Mandic R, Peterlin BM: **Interaction between Nef and phosphatidylinositol-3-kinase leads to activation of p21-activated kinase and increased production of HIV**. *Virology* 2002, **294**(2):246-255.

77. Li PL, Wang T, Buckley KA, Chenine AL, Popov S, Ruprecht RM: **Phosphorylation of HIV Nef by cAMP-dependent protein kinase**. *Virology* 2005, **331**(2):367-374.

78. Coates K, Harris M: **The human immunodeficiency virus type 1 Nef protein functions as a protein kinase C substrate in vitro**. *J Gen Virol* 1995, **76 (Pt 4)**:837-844.

79. Trible RP, Emert-Sedlak L, Smithgall TE: **HIV-1 Nef selectively activates Src family kinases Hck, Lyn, and c-Src through direct SH3 domain interaction**. *J Biol Chem* 2006, **281**(37):27029-27038.

80. Greenway AL, McPhee DA, Allen K, Johnstone R, Holloway G, Mills J, Azad A, Sankovich S, Lambert P: **Human immunodeficiency virus type 1 Nef binds to tumor suppressor p53 and protects cells against p53-mediated apoptosis**. *J Virol* 2002, **76**(6):2692-2702.

81. Meggio F, D'Agostino DM, Ciminale V, Chieco-Bianchi L, Pinna LA: **Phosphorylation of HIV-1 Rev protein: implication of protein kinase CK2 and pro-directed kinases**. *Biochem Biophys Res Commun* 1996, **226**(2):547-554.

82. Vendel AC, Lumb KJ: **Molecular recognition of the human coactivator CBP by the HIV-1 transcriptional activator Tat**. *Biochemistry* 2003, **42**(4):910-916.

83. Deng L, de la Fuente C, Fu P, Wang L, Donnelly R, Wade JD, Lambert P, Li H, Lee CG, Kashanchi F: **Acetylation of HIV-1 Tat by CBP/P300 increases transcription of integrated HIV-1 genome and enhances binding to core histones**. *Virology* 2000, **277**(2):278-295.

84. Holmes AM: **In vitro phosphorylation of human immunodeficiency virus type 1 Tat protein by protein kinase C: evidence for the phosphorylation of amino acid residue serine-46**. *Arch Biochem Biophys* 1996, **335**(1):8-12.

85. Yang X, Gabuzda D: **Mitogen-activated protein kinase phosphorylates and regulates the HIV-1 Vif protein**. *J Biol Chem* 1998, **273**(45):29879-29887.

86. Kino T, Gragerov A, Slobodskaya O, Tsopanomichalou M, Chrousos GP, Pavlakis GN: **Human immunodeficiency virus type 1 (HIV-1) accessory protein Vpr induces transcription of the HIV-1 and glucocorticoid-responsive promoters by binding directly to p300/CBP coactivators**. *J Virol* 2002, **76**(19):9724-9734.

87. Friborg J, Ladha A, Gottlinger H, Haseltine WA, Cohen EA: **Functional analysis of the phosphorylation sites on the human immunodeficiency virus type 1 Vpu protein**. *J Acquir Immune Defic Syndr Hum Retrovirol* 1995, **8**(1):10-22.

88. Chen SS, Yang P, Ke PY, Li HF, Chan WE, Chang DK, Chuang CK, Tsai Y, Huang SC: **Identification of the LWYIK motif located in the human immunodeficiency virus type 1 transmembrane gp41 protein as a distinct determinant for viral infection**. *J Virol* 2009, **83**(2):870-883.

89. Tahirov TH, Babayeva ND, Varzavand K, Cooper JJ, Sedore SC, Price DH: **Crystal structure of HIV-1 Tat complexed with human P-TEFb**. *Nature* 2010, **465**(7299):747-751.

90. Tastan O, Klein-Seetharaman J, Meirovitch H: **The effect of loops on the structural organization of alpha-helical membrane proteins**. *Biophys J* 2009, **96**(6):2299-2312.

91. Dunker AK, Brown CJ, Lawson JD, Iakoucheva LM, Obradovic Z: **Intrinsic disorder and protein function**. *Biochemistry* 2002, **41**(21):6573-6582.

92. Kim B, Ayran JC, Sagar SG, Adman ET, Fuller SM, Tran NH, Horrigan J: **New human immunodeficiency virus, type 1 reverse transcriptase (HIV-1 RT) mutants with increased fidelity of DNA synthesis. Accuracy, template binding, and processivity**. *J Biol Chem* 1999, **274**(39):27666-27673.

93. Vilar M, Sauri A, Marcos JF, Mingarro I, Perez-Paya E: **Transient structural ordering of the RNA-binding domain of carnation mottle virus p7 movement protein modulates nucleic acid binding**. *Chembiochem* 2005, **6**(8):1391-1396.

94. Chen Y, Xu D: **Computational analyses of high-throughput protein-protein interaction data**. *Curr Protein Pept Sci* 2003, **4**(3):159-181.

95. Mohan A, Oldfield CJ, Radivojac P, Vacic V, Cortese MS, Dunker AK, Uversky VN: **Analysis of molecular recognition features (MoRFs)**. *J Mol Biol* 2006, **362**(5):1043-1059.

96. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL *et al*: **The Pfam protein families database**. *Nucleic Acids Res* 2004, **32**(Database issue):D138-141.

97. Sigrist CJ, Cerutti L, de Castro E, Langendijk-Genevaux PS, Bulliard V, Bairoch A, Hulo N: **PROSITE, a protein domain database for functional characterization and annotation**. *Nucleic Acids Res* 2010, **38**(Database issue):D161-166.

98. Arold ST, Baur AS: **Dynamic Nef and Nef dynamics: how structure could explain the complex activities of this small HIV protein**. *Trends Biochem Sci* 2001, **26**(6):356-363.

99. Geyer M, Peterlin BM: **Domain assembly, surface accessibility and sequence conservation in full length HIV-1 Nef**. *FEBS Lett* 2001, **496**(2-3):91-95.

100. Arhel NJ, Kirchhoff F: **Implications of Nef: host cell interactions in viral persistence and progression to AIDS**. *Curr Top Microbiol Immunol* 2009, **339**:147-175.

101. Liu X, Schrager JA, Lange GD, Marsh JW: **HIV Nef-mediated cellular phenotypes are differentially expressed as a function of intracellular Nef concentrations**. *J Biol Chem* 2001, **276**(35):32763-32770.

102. Agopian K, Wei BL, Garcia JV, Gabuzda D: **CD4 and MHC-I downregulation are conserved in primary HIV-1 Nef alleles from brain and lymphoid tissues, but Pak2 activation is highly variable**. *Virology* 2007, **358**(1):119-135.

103. Miller MD, Warmerdam MT, Ferrell SS, Benitez R, Greene WC: **Intravirion generation of the C-terminal core domain of HIV-1 Nef by the HIV-1 protease is insufficient to enhance viral infectivity**. *Virology* 1997, **234**(2):215-225.

104. Huang da W, Sherman BT, Lempicki RA: **Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources**. *Nat Protoc* 2009, **4**(1):44-57.

105. Edwards RJ, Davey NE, Shields DC: **CompariMotif: quick and easy comparisons of sequence motifs**. *Bioinformatics* 2008, **24**(10):1307-1309.

106. Deacon NJ, Tsykin A, Solomon A, Smith K, Ludford-Menting M, Hooker DJ, McPhee DA, Greenway AL, Ellett A, Chatfield C *et al*: **Genomic structure of an attenuated quasi species of HIV-1 from a blood transfusion donor and recipients**. *Science* 1995, **270**(5238):988-991.

107. Kestler HW, 3rd, Ringler DJ, Mori K, Panicali DL, Sehgal PK, Daniel MD, Desrosiers RC: **Importance of the nef gene for maintenance of high virus loads and for development of AIDS**. *Cell* 1991, **65**(4):651-662.

108. Robert-Guroff M, Popovic M, Gartner S, Markham P, Gallo RC, Reitz MS: **Structure and expression of tat-, rev-, and nef-specific transcripts of human immunodeficiency virus type 1 in infected lymphocytes and macrophages**. *J Virol* 1990, **64**(7):3391-3398.

109. Kienzle N, Bachmann M, Muller WE, Muller-Lantzsch N: **Expression and cellular localization of the Nef protein from human immunodeficiency virus-1 in stably transfected B-cells**. *Arch Virol* 1992, **124**(1-2):123-132.

110. Macreadie IG, Ward AC, Failla P, Grgacic E, McPhee D, Azad AA: **Expression of HIV-1 nef in yeast: the 27 kDa Nef protein is myristylated and fractionates with the nucleus**. *Yeast* 1993, **9**(6):565-573.

111. Murti KG, Brown PS, Ratner L, Garcia JV: **Highly localized tracks of human immunodeficiency virus type 1 Nef in the nucleus of cells of a human CD4+ T-cell line**. *Proc Natl Acad Sci U S A* 1993, **90**(24):11895-11899.

112. Ranki A, Lagerstedt A, Ovod V, Aavik E, Krohn KJ: **Expression kinetics and subcellular localization of HIV-1 regulatory proteins Nef, Tat and Rev in acutely and chronically infected lymphoid cell lines**. *Arch Virol* 1994, **139**(3-4):365-378.

113. Greenway AL, Holloway G, McPhee DA: **HIV-1 Nef: a critical factor in viral-induced pathogenesis**. *Adv Pharmacol* 2000, **48**:299-343.

114. Collette Y, Dutartre H, Benziane A, Ramos M, Benarous R, Harris M, Olive D: **Physical and functional interaction of Nef with Lck. HIV-1 Nef-induced T-cell signaling defects**. *J Biol Chem* 1996, **271**(11):6333-6341.

115. Koenig S, Fuerst TR, Wood LV, Woods RM, Suzich JA, Jones GM, de la Cruz VF, Davey RT, Jr., Venkatesan S, Moss B *et al*: **Mapping the fine specificity of a cytolytic T cell response to HIV-1 nef protein**. *J Immunol* 1990, **145**(1):127-135.

116. Jin YJ, Zhang X, Cai CY, Burakoff SJ: **Alkylating HIV-1 Nef - a potential way of HIV intervention**. *AIDS Res Ther* 2010, **7**:26.

117. Petsalaki E, Russell RB: **Peptide-mediated interactions in biological systems: new discoveries and applications**. *Curr Opin Biotechnol* 2008, **19**(4):344-350.

118. Pawson T, Scott JD: **Signaling through scaffold, anchoring, and adaptor proteins**. *Science* 1997, **278**(5346):2075-2080.

119. Doherty CP: **Host-pathogen interactions: the role of iron**. *J Nutr* 2007, **137**(5):1341-1344.

120. Weiss G: **Iron and immunity: a double-edged sword**. *Eur J Clin Invest* 2002, **32 Suppl 1**:70-78.

121. Drakesmith H, Prentice A: **Viral infection and iron metabolism**. *Nat Rev Microbiol* 2008, **6**(7):541-552.

122. Weiss G, Wachter H, Fuchs D: **Linkage of cell-mediated immunity to iron metabolism**. *Immunol Today* 1995, **16**(10):495-500.

123. Seligman PA, Kovar J, Gelfand EW: **Lymphocyte proliferation is controlled by both iron availability and regulation of iron uptake pathways**. *Pathobiology* 1992, **60**(1):19-26.

124. De Sousa M: **T lymphocytes and iron overload: novel correlations of possible significance to the biology of the immunological system**. *Mem Inst Oswaldo Cruz* 1992, **87 Suppl 5**:23-29.

125. Brekelmans P, van Soest P, Leenen PJ, van Ewijk W: **Inhibition of proliferation and differentiation during early T cell development by anti-transferrin receptor antibody**. *Eur J Immunol* 1994, **24**(11):2896-2902.

126. Savarino A, Pescarmona GP, Boelaert JR: **Iron metabolism and HIV infection: reciprocal interactions with potentially harmful consequences?** *Cell Biochem Funct* 1999, **17**(4):279-287.

127. Boelaert JR, Weinberg GA, Weinberg ED: **Altered iron metabolism in HIV infection: mechanisms, possible consequences, and proposals for management**. *Infect Agents Dis* 1996, **5**(1):36-46.

128. de Monye C, Karcher DS, Boelaert JR, Gordeuk VR: **Bone marrow macrophage iron grade and survival of HIV-seropositive patients**. *AIDS* 1999, **13**(3):375-380.

129. Gordeuk VR, Delanghe JR, Langlois MR, Boelaert JR: **Iron status and the outcome of HIV infection: an overview**. *J Clin Virol* 2001, **20**(3):111-115.

130. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT *et al*: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium**. *Nat Genet* 2000, **25**(1):25-29.

131. **The Universal Protein Resource (UniProt) in 2010**. *Nucleic Acids Res* 2010, **38**(Database issue):D142-148.

132. Funahashi AM, Y.; Jouraku, A.; Morohashi, M.; Kikuchi, N.; Kitano, H.: **CellDesigner 3.5: A Versatile Modeling Tool for Biochemical Networks**. *Proceedings of the IEEE* 2008, **96**(8):1254.

133. Kurtoglu E, Ugur A, Baltaci AK, Mogolkoc R, Undar L: **Activity of neutrophil NADPH oxidase in iron-deficient anemia**. *Biol Trace Elem Res* 2003, **96**(1-3):109-115.

134. Olivetta E, Mallozzi C, Ruggieri V, Pietraforte D, Federico M, Sanchez M: **HIV-1 Nef induces p47(phox) phosphorylation leading to a rapid superoxide anion release from the U937 human monoblastic cell line**. *J Cell Biochem* 2009, **106**(5):812-822.

135. Olivetta E, Pietraforte D, Schiavoni I, Minetti M, Federico M, Sanchez M: **HIV-1 Nef regulates the release of superoxide anions from human macrophages**. *Biochem J* 2005, **390**(Pt 2):591-602.

136. Lachgar A, Sojic N, Arbault S, Bruce D, Sarasin A, Amatore C, Bizzini B, Zagury D, Vuillaume M: **Amplification of the inflammatory cellular redox state by**

human immunodeficiency virus type 1-immunosuppressive tat and gp160 proteins. *J Virol* 1999, **73**(2):1447-1452.

137. Jana A, Pahan K: **Human immunodeficiency virus type 1 gp120 induces apoptosis in human primary neurons through redox-regulated activation of neutral sphingomyelinase**. *J Neurosci* 2004, **24**(43):9531-9540.

138. Nong Y, Kandil O, Tobin EH, Rose RM, Remold HG: **The HIV core protein p24 inhibits interferon-gamma-induced increase of HLA-DR and cytochrome b heavy chain mRNA levels in the human monocyte-like cell line THP1**. *Cell Immunol* 1991, **132**(1):10-16.

139. Lake JA, Carr J, Feng F, Mundy L, Burrell C, Li P: **The role of Vif during HIV-1 infection: interaction with novel host cellular factors**. *J Clin Virol* 2003, **26**(2):143-152.

140. Cullen BR: **HIV-1 auxiliary proteins: making connections in a dying cell**. *Cell* 1998, **93**(5):685-692.

141. Dooher JE, Schneider BL, Reed JC, Lingappa JR: **Host ABCE1 is at plasma membrane HIV assembly sites and its dissociation from Gag is linked to subsequent events of virus production**. *Traffic* 2007, **8**(3):195-211.

142. Rodnina MV: **Protein synthesis meets ABC ATPases: new roles for Rli1/ABCE1**. *EMBO Rep* 2010, **11**(3):143-144.

143. Armitage AE, McMichael AJ, Drakesmith H: **Reflecting on a quarter century of HIV research**. *Nat Immunol* 2008, **9**(8):823-826.

144. Valiathan RR, Resh MD: **Expression of human immunodeficiency virus type 1 gag modulates ligand-induced downregulation of EGF receptor**. *J Virol* 2004, **78**(22):12386-12394.

145. Lob S, Konigsrainer A: **Is IDO a key enzyme bridging the gap between tumor escape and tolerance induction?** *Langenbecks Arch Surg* 2008, **393**(6):995-1003.

146. Austin CJ, Mailu BM, Maghzal GJ, Sanchez-Perez A, Rahlfs S, Zocher K, Yuasa HJ, Arthur JW, Becker K, Stocker R *et al*: **Biochemical characteristics and inhibitor selectivity of mouse indoleamine 2,3-dioxygenase-2**. *Amino Acids* 2010.

147. Boasso A, Herbeuval JP, Hardy AW, Anderson SA, Dolan MJ, Fuchs D, Shearer GM: **HIV inhibits CD4+ T-cell proliferation by inducing indoleamine 2,3-dioxygenase in plasmacytoid dendritic cells**. *Blood* 2007, **109**(8):3351-3359.

148. Mirani M, Elenkov I, Volpi S, Hiroi N, Chrousos GP, Kino T: **HIV-1 protein Vpr suppresses IL-12 production from human monocytes by enhancing glucocorticoid action: potential implications of Vpr coactivator activity for the innate and cellular immunity deficits observed in HIV-1 infection**. *J Immunol* 2002, **169**(11):6361-6368.

149. Zheng YH, Plemenitas A, Fielding CJ, Peterlin BM: **Nef increases the synthesis of and transports cholesterol to lipid rafts and HIV-1 progeny virions**. *Proc Natl Acad Sci U S A* 2003, **100**(14):8460-8465.

150. Vazquez JA, Skiest DJ, Tissot-Dupont H, Lennox JL, Boparai N, Isaacs R: **Safety and efficacy of posaconazole in the long-term treatment of azole-refractory oropharyngeal and esophageal candidiasis in patients with HIV infection**. *HIV Clin Trials* 2007, **8**(2):86-97.

151. Rupp B, Raub S, Marian C, Holtje HD: **Molecular design of two sterol 14alpha-demethylase homology models and their interactions with the azole antifungals ketoconazole and bifonazole**. *J Comput Aided Mol Des* 2005, **19**(3):149-163.

152. Feisst C, Pergola C, Rakonjac M, Rossi A, Koeberle A, Dodt G, Hoffmann M, Hoernig C, Fischer L, Steinhilber D *et al*: **Hyperforin is a novel type of 5-lipoxygenase inhibitor with high efficacy in vivo**. *Cell Mol Life Sci* 2009, **66**(16):2759-2771.

153. Radmark O, Werz O, Steinhilber D, Samuelsson B: **5-Lipoxygenase: regulation of expression and enzyme activity**. *Trends Biochem Sci* 2007, **32**(7):332-341.

154. Maccarrone M, Navarra M, Catani V, Corasaniti MT, Bagetta G, Finazzi-Agro A: **Cholesterol-dependent modulation of the toxicity of HIV-1 coat protein gp120 in human neuroblastoma cells**. *J Neurochem* 2002, **82**(6):1444-1452.

155. Maccarrone M, Navarra M, Corasaniti MT, Nistico G, Finazzi Agro A: **Cytotoxic effect of HIV-1 coat glycoprotein gp120 on human neuroblastoma CHP100 cells involves activation of the arachidonate cascade**. *Biochem J* 1998, **333 (Pt 1)**:45-49.

156. Coffey MJ, Phare SM, Cinti S, Peters-Golden M, Kazanjian PH: **Granulocyte-macrophage colony-stimulating factor upregulates reduced 5-lipoxygenase metabolism in peripheral blood monocytes and neutrophils in acquired immunodeficiency syndrome**. *Blood* 1999, **94**(11):3897-3905.

157. Coffey MJ, Phare SM, Kazanjian PH, Peters-Golden M: **5-Lipoxygenase metabolism in alveolar macrophages from subjects infected with the human immunodeficiency virus**. *J Immunol* 1996, **157**(1):393-399.

158. Coffey MJ, Phare SM, George S, Peters-Golden M, Kazanjian PH: **Granulocyte colony-stimulating factor administration to HIV-infected subjects augments reduced leukotriene synthesis and anticryptococcal activity in neutrophils**. *J Clin Invest* 1998, **102**(4):663-670.

Appendices

Appendix A: List of motifs defining hotspots per HIV protein accessible at

http://bioinformatics.biomed.drexel.edu/mahdi/

This file is a nine tab Excel spread sheet containing motifs shared by HIV proteins and some of the neighbors of HIV protein targeted hub proteins. Each tab lists motifs for an HIV protein with its corresponding details. Headings *Hub ID* and *Hub Symbol* represent the Entrez ID and gene symbol of the hub protein to which the motif belongs. *Pattern* is the regular expression of the motif. *Info Content* is the information content of the motif pattern. The *p value* is computed by statistical enrichment of the motif among neighbors of the hub protein in comparison to HPRD proteins. The number of neighbors of a hub protein and the neighbors on which the motif is present are shown with the symbols *# of H2s* and *H2s w/Motif*, respectively. *Start* and *End* headings refer to the start and end positions of the motif on the corresponding HIV protein sequence, calculated based on the most common positions observed on the HIV protein sequences.

Vita

Mahdi Sarmady ms872@drexel.edu , 484-416-0005

Education
Drexel University, Philadelphia, PA
PhD in Biomedical Engineering, September 2010

University of Tehran, Tehran, Iran
MSc in Computer Engineering, July 2007

University of Tehran, Tehran, Iran
BSc in Computer Engineering, July 2005

Research
Bioinformatics Research Assistant, 2008-2010
Center for Integrated Bioinformatics
Drexel University, Philadelphia, PA

Formal Methods Laboratory, 2005-2007
University of Tehran, Tehran, Iran

Publications
Sarmady M, Bukrinsky M, Tozeren A, HIV-1 sequence hotspots for crosstalk with host hub proteins (in preparation).

Sarmady M, Tozeren A, HIV-1 Nef expresses sequence hotspots for binding to host proteins (in preparation).

Sarmady M, Dawany N, Tozeren A, Connectivity map for viral crosstalk with host iron binding proteins (in preparation).

Teaching Experience
Teaching Assistant, 2008-2010
Drexel University, Philadelphia, PA

- Computational Bioengineering
- Biomedical Ethics and Law
- Medical Device Development